



UNIL | Université de Lausanne



Genes: from sequence to function

# We got the sequence: now what?

# About me

Carlo Rivolta, Ph.D.

Molecular geneticist working on hereditary diseases

Department of Medical Genetics (DGM)

University of Lausanne

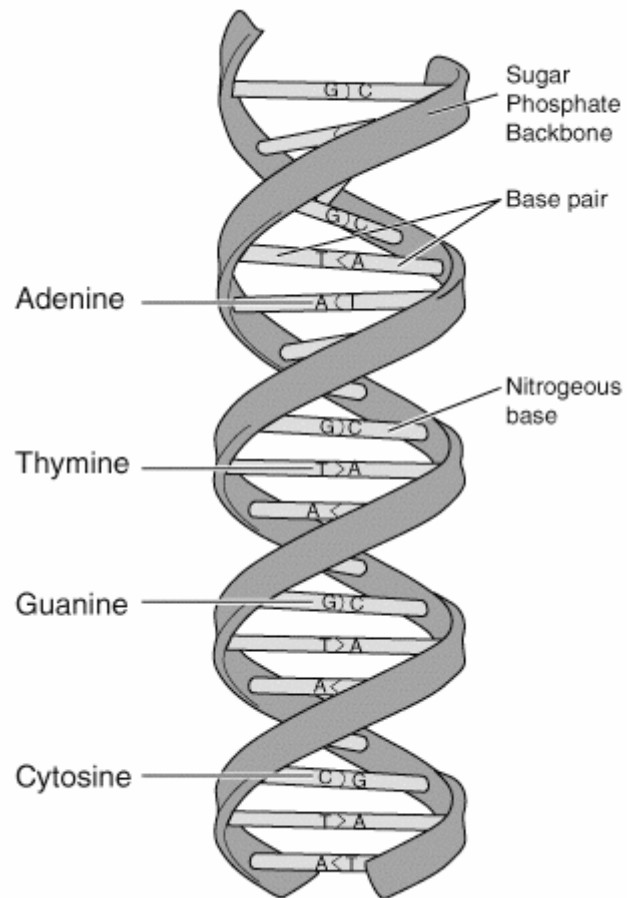
[www.unil.ch/dgm](http://www.unil.ch/dgm)

carlo.rivolta@unil.ch

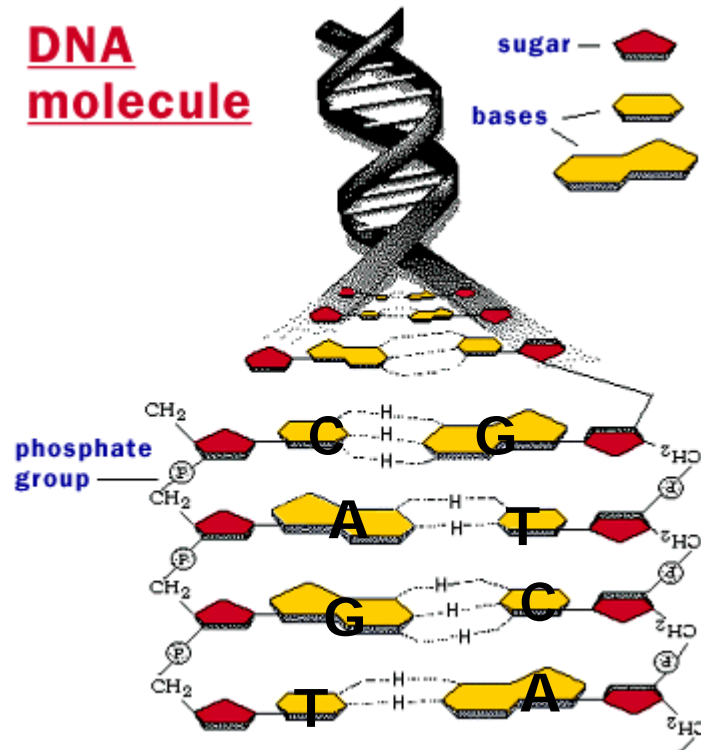
# Course plan

- Sept 17 (10:00-12:00)
  - Gene identification strategies
- Oct 5 (13:00-17:00) **SALLE 2019 GENOPODE**
  - Gene identification strategies II, Sequence databases and genome browsers
- Oct 26 (13:00-15:00) **SALLE 2019 GENOPODE**
  - Exercises

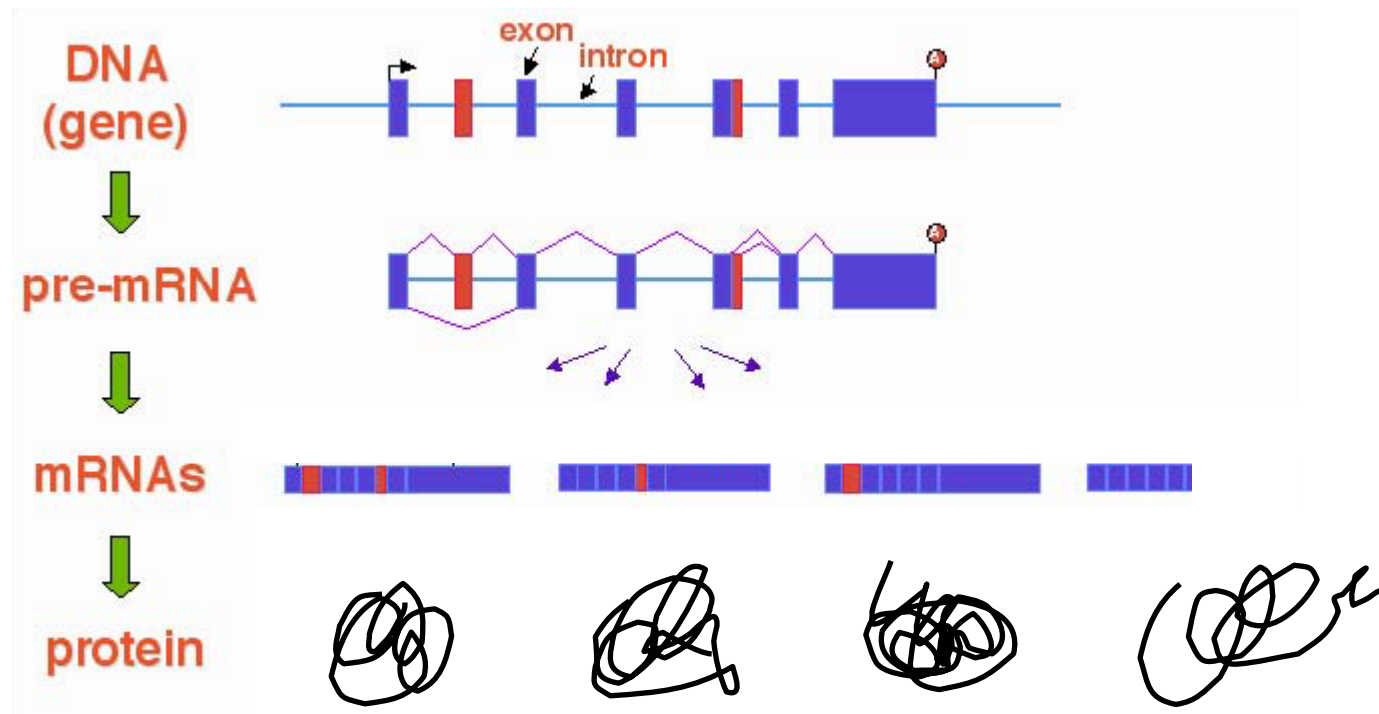
# Prerequisites



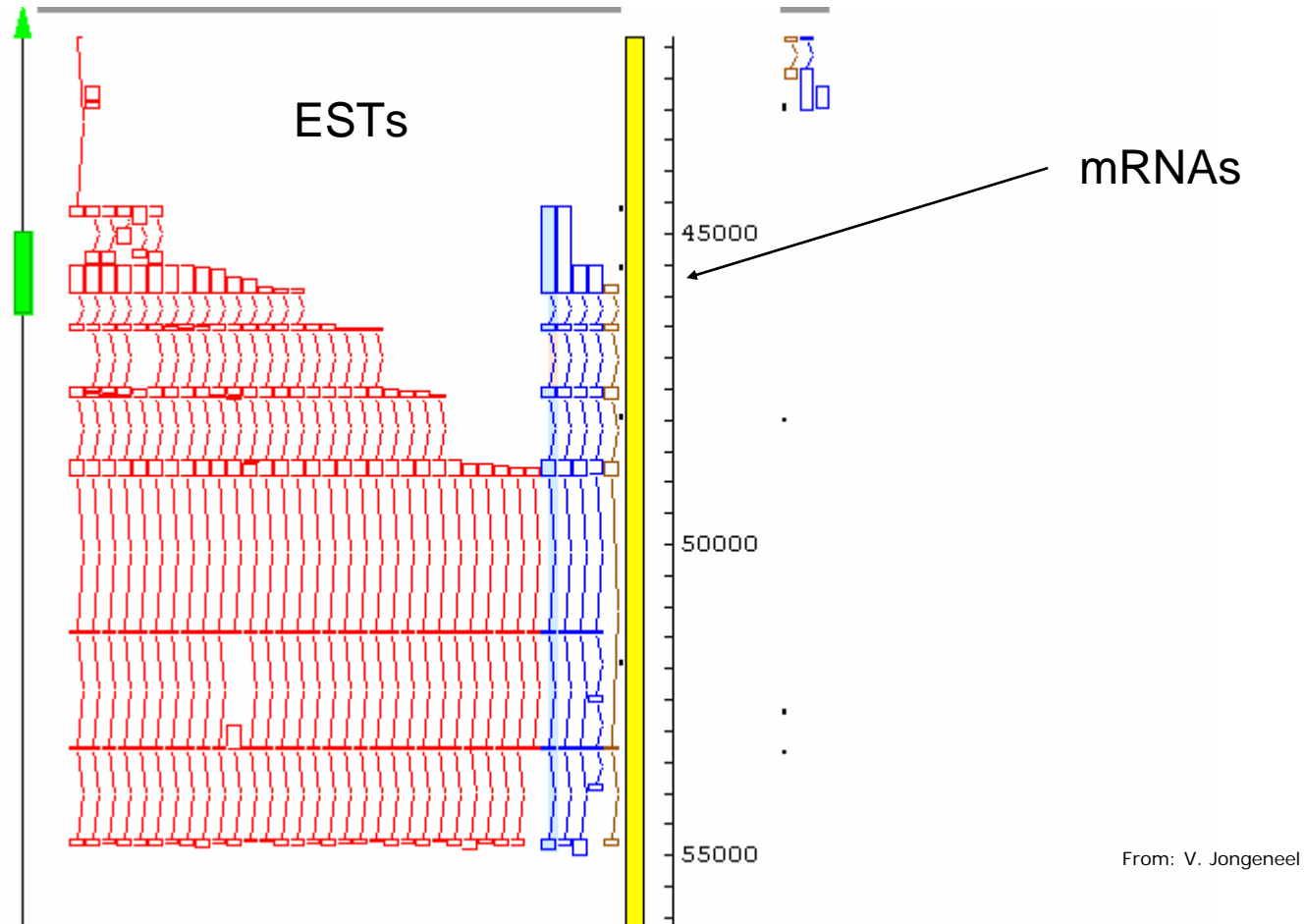
G: C  
A: T



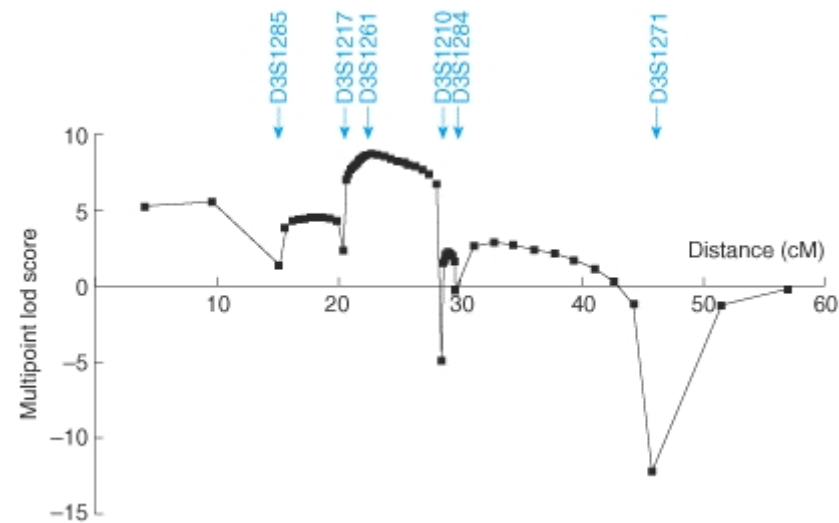
# Prerequisites



# Prerequisites



# Prerequisites



From: Strachan and Read  
Human Molecular Genetics

# Prerequisites

The screenshot shows the NCBI homepage in a Windows Internet Explorer browser window. The address bar displays <http://www.ncbi.nlm.nih.gov/>. The browser's menu bar includes File, Edit, View, Favorites, Tools, and Help. The toolbar shows icons for Google, NCBI, Human (H...), Webmail, UNIL - An..., UNIL UNL ens..., UNIL Introducti..., UNIL FBM UNIL..., UNIL DGM UNIL..., and NCBI. The main content area features the NCBI logo and the text "National Center for Biotechnology Information", "National Library of Medicine", and "National Institutes of Health". A navigation bar includes links to PubMed, All Databases, BLAST, OMIM, Books, TaxBrowser, and Structure. Below this is a search bar with a dropdown menu set to "All Databases" and a "Go" button. The left sidebar contains a "SITE MAP" with links to "Alphabetical List" and "Resource Guide", "About NCBI" with an "Introduction to NCBI" link, "GenBank" with "Sequence submission support and software", "Literature databases" with links to "PubMed, OMIM, Books, and PubMed Central", "Molecular databases" with links to "Sequences, structures, and taxonomy", and "Genomic biology" with links to "The human genome, whole genomes, and related resources". The main content area has a "What does NCBI do?" section with a paragraph about its establishment in 1988 and a "More..." link. Below this is a "GenBank vs. RefSeq" section with a paragraph about the distinctions between GenBank, RefSeq, TPA, and UniProt, and a "Click here" link. Further down is a "New dbGaP" section with a paragraph about the database of Genotype and Phenotype (GWA) studies and a "Click here" link. On the right, a "Hot Spots" section lists various resources: Assembly Archive, Clusters of orthologous groups, Coffee Break, Genes & Disease, NCBI Handbook, Electronic PCR, Entrez Home, Entrez Tools, Gene expression omnibus (GEO), Human genome resources, Influenza Virus Resource, Map Viewer, dbMHC, and Mouse genome resources.

NCBI HomePage - Windows Internet Explorer  
http://www.ncbi.nlm.nih.gov/  
File Edit View Favorites Tools Help  
Google NCBI ... Human (H...) Webmail ... UNIL - An... UNIL UNL ens... UNIL Introducti... UNIL FBM UNIL... UNIL DGM UNIL... NCBI

NCBI  
National Center for Biotechnology Information  
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search All Databases for Go

**SITE MAP**  
Alphabetical List  
Resource Guide

**About NCBI**  
An introduction to NCBI

**GenBank**  
Sequence submission support and software

**Literature databases**  
PubMed, OMIM, Books, and PubMed Central

**Molecular databases**  
Sequences, structures, and taxonomy

**Genomic biology**  
The human genome, whole genomes, and related resources

**What does NCBI do?**  
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)

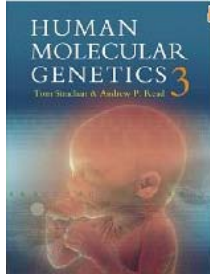
**GenBank vs. RefSeq**  
Confused about the distinctions between GenBank, RefSeq, TPA and UniProt? [Click here](#) for a brief description of the databases and their differences.

**New dbGaP**  
NCBI's dbGaP Genome Wide Association Database  
NCBI's dbGaP (database of Genotype and Phenotype) provides data from Genome Wide Association (GWA) studies. The resource is intended to help elucidate the link between genes and disease. For each study, users have access to detailed information about the phenotypic variables measured and pre-computed associations between subjects' phenotypes and genotypes. [Click here](#) to read the [press release](#). To read more about GWA projects, see NCBI's [GWA resource page](#).

**Hot Spots**  
► Assembly Archive  
► Clusters of orthologous groups  
► Coffee Break, Genes & Disease, NCBI Handbook  
► Electronic PCR  
► Entrez Home  
► Entrez Tools  
► Gene expression omnibus (GEO)  
► Human genome resources  
► Influenza Virus Resource  
► Map Viewer  
► dbMHC  
► Mouse genome resources



# Suggested readings

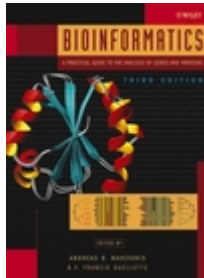


## **Human Molecular Genetics, 3rd Edition (2003)**

Tom Strachan

Andrew Read

ISBN: 978-0815341826

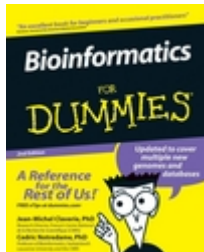


## **Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition (2005)**

Andreas D. Baxevanis

Francis Ouellette

ISBN: 978-0-471-47878-2



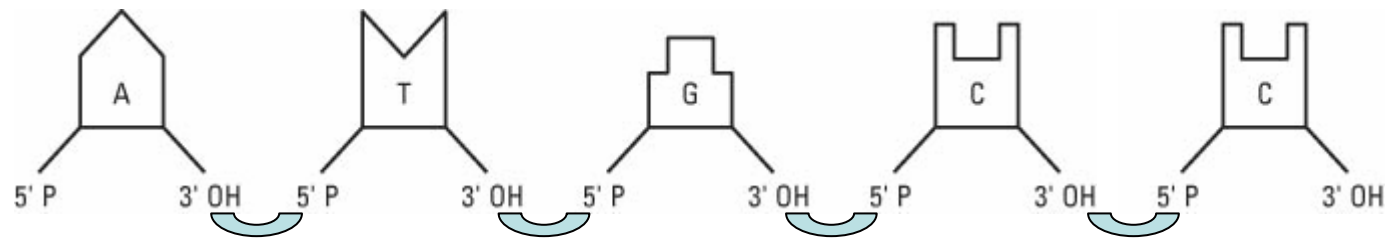
## **Bioinformatics For Dummies, 2nd Edition (2006)**

Jean-Michel Claverie

Cedric Notredame

ISBN: 978-0-470-08985-9

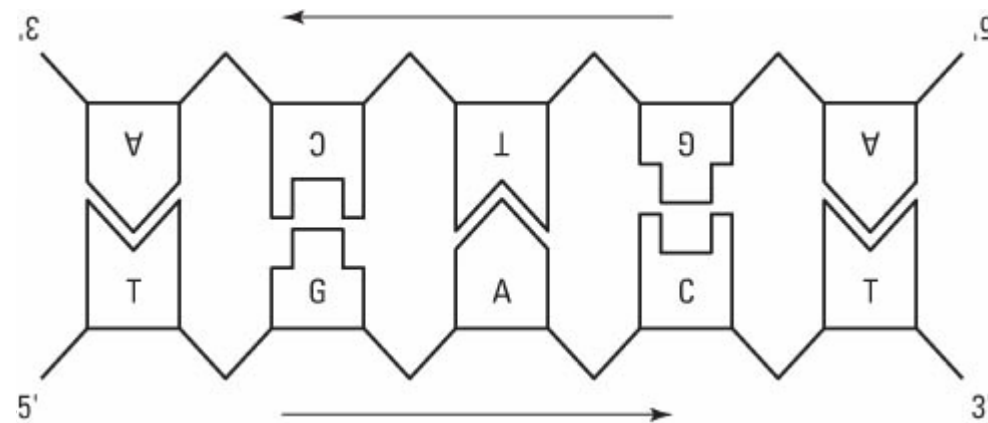
# The basics



5'-ATGCC-3'

ATGCC

# The basics



TGACT  
or  
AGTCA

**ALWAYS READ 5' TO 3'**

# The basics

3'-ENNASUAL-5'

# The basics

```
GCCCTGCGGAGCAAGGTACCTCACACTTCATGAGCGAGTTAAGATGGGTTTCACAATTTT
TCAAGCAAGGAAACGGGCTCGGAGGTCTTGAACACCTGCTACCCAATAGCAGAACAGCTA
CTGGAACATAAATCCTCTGATTTCAAATAACAGCCCCGCCACTACCACTAAGTGAAGTC
ATCCACAACACACACCGGACCACTCTAAGCTTTTGAAGATCGGCTCGCTTTGGGGAACA
GGTCTTGAGAGAACATCCCTTTTAAGGTCAGAACAAAGGTATTTATAGGTCCAGGTCG
TGTCGCCGAGGCGCCACCAACATGAGCTGGAGCAAAAAGAAAGGGATGGGGGACTTG
GAGTAGGCATAGGGGCGGCCCCCTCAAGCAGGGTGGCTGGGACTCTTAAGGGTCAGCGA
GAAGAGAACACACACTCCAGCTCCCGCTTTATTCGGTCAGATACTGACGGTTGGGATGCC
TGACAAGGAATTTCTTTTCGCCACACTGAGAAATACCCGACGCGGCCACCCAGGCCTGA
CTTCGGGTGGTGGCTGTGCTGCGTGTGCGGTACGCGCTCACGTGGCCAGCGCGGCTT
GTGGCGGAGCTTCTGAAACTAGGCGGCGAGGCGGAGCGCTGTGGCACTGCTGCGCT
CTGCTGCGCTCGGGTGTCTTTTGGCGCGTGGGTGCGCGCGGGGAGAAGCGTGAGGGGA
CAGATTTGTGACCGGCGCGGTTTGTGCTAGCTTACTCGGCCAAAAAGAACTGCACCTC
TGGAGCGGGTTAGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGT
CCAGCGTGGCGGGGAGCGCTCACGCCCCGGGTGCTGCGCGGCTTCTTGGCCTTTTG
TCTCTGCCAACCCCCACCATGCCCTGAGAGAAAGTCTTGGCCGAAGGCAGATTTTCGC
CAAGCAAAATTCGAGCCCCCGCTTCCCTGGGTCTCCATTTCCCGCTTCCGCGCGGCTT
TTGGGCTCCGCTTTCAGCTCAAGACTTAACCTCCCTCCAGCTGTCCAGATGACGCCAT
CTGAAATTTCTTGGAAACACGATCACTTTAACGGAATATTGCTGTTTTGGGGAAGTGT
TACAGCTGCTGGGCACGCTGATTTGCCCTTACTTAAGCCCTGGTAATTGCTGATTTCCG
AAGACATGCTGATGGGAATTACCAGGCGGCGTGGTCTCTAAGTGGAGCCCTCTGTCCCC
ACTAGCCACGCTCACTGGTTAGCGTGATTGAACTAAATCGTATGAAATCCTCTTCTC
TAGTCGCACTAGCCACGTTTCGAGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCA
CAGCTGTAAATGTTCCCATCTCACAGTAAGCTGTACCCTTCCAGGAGATGGGACTGA
ATTAGAATTCAAACAAATTTCCAGCGCTTCTGAGTTTACCTCAGTCACATAATAAGGA
ATGCATCCCTGTGTAAGTGCAATTTGGTCTTCTGTTTTGCGAGACTTATTTACCAAGCAT
GGAGGAATATCGTAGGTAAAAATGCCATTTGGATCCAAAGAGAGGCCAACATTTTGGAA
ATTTTTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTTTATATACTTTTATAAAT
ACACCGAGAAAGTGTCTTAAATAATGCTTGCTAAAAACCCAGTACGTACAGTGTGTC
TTAGAACCATAAACTGTCTCTTATGTGTGTATAAATCCAGTTAAACACATAATCATCGTT
TGCAGGTTAACCACATGATAAATATAGAACGCTCTAGTGGATAAAGAGGAACTGGCCCC
TGACTAGCAGTAGGAACAATTAACAAATCAGAAGCATTAATGTTACTTTTATGGCAGA
AGTTGTCCAATTTTGGTTTCAGTACTCCTTATACTCTTAAAAATGATCTAGGACCCCC
GGAGTGTCTTTGTTTATGTAGCTTACCATATTAGAAATTTAAACATAAGAAATTTAAGGT
GGGCGTGGTGGCTCACGCCGTGAATCCAGCACTTTGGGAGGCGGAGGTGGGCGGATCAC
TTGAGGCCAGAAAGTTTGAACACGCTGGCCACATGTTGAAACCTATCTCTACTAAAA
ATACAAAAATGTGCTGCGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGG
GAGGCAGGAGAAATCGCTTGAACCTGGAGGCAGAGGTTGCAGTGAGCCAAAGATCATGCCA
CTGCACCTAGCCTGGGCCACATAGCATGACTCTGTCTCAAAACAAACAAACAAACAA
AACTAAGAAATTTAAAGTTAATTTACTTAAAAATAATGAAAGCTAACCCATTGCAATATTAT
CACAACTTCTTAGGAAAAATAACTTTTGAACAAAGTGAGTGGAATAGTTTTTACATTT
TTTGAGTCTCTTTAAATGCTGGCTAAATAGAGATAGCTGGATTCACTTATCTGTGTCT
AATCTGTTATTTTGTAGAAATGATGTGAAAAAAATTAACCTCACGTTGAAAAAAGGAAT
ATTTTAATAGTTTTTCACTTACTTTTGGTATTTTTCCTTGTACTTTGTCATAGATTTTCA
AAGATCTAATAGATATACCATAGGTCTTCCCATGTGCAACATCATGCAGTGATTTATTT
GGAAGATAGTGGTGTCTGAATTTATACAAAGTTTCCAAATATTGATAAATTGCATTAAAC
```

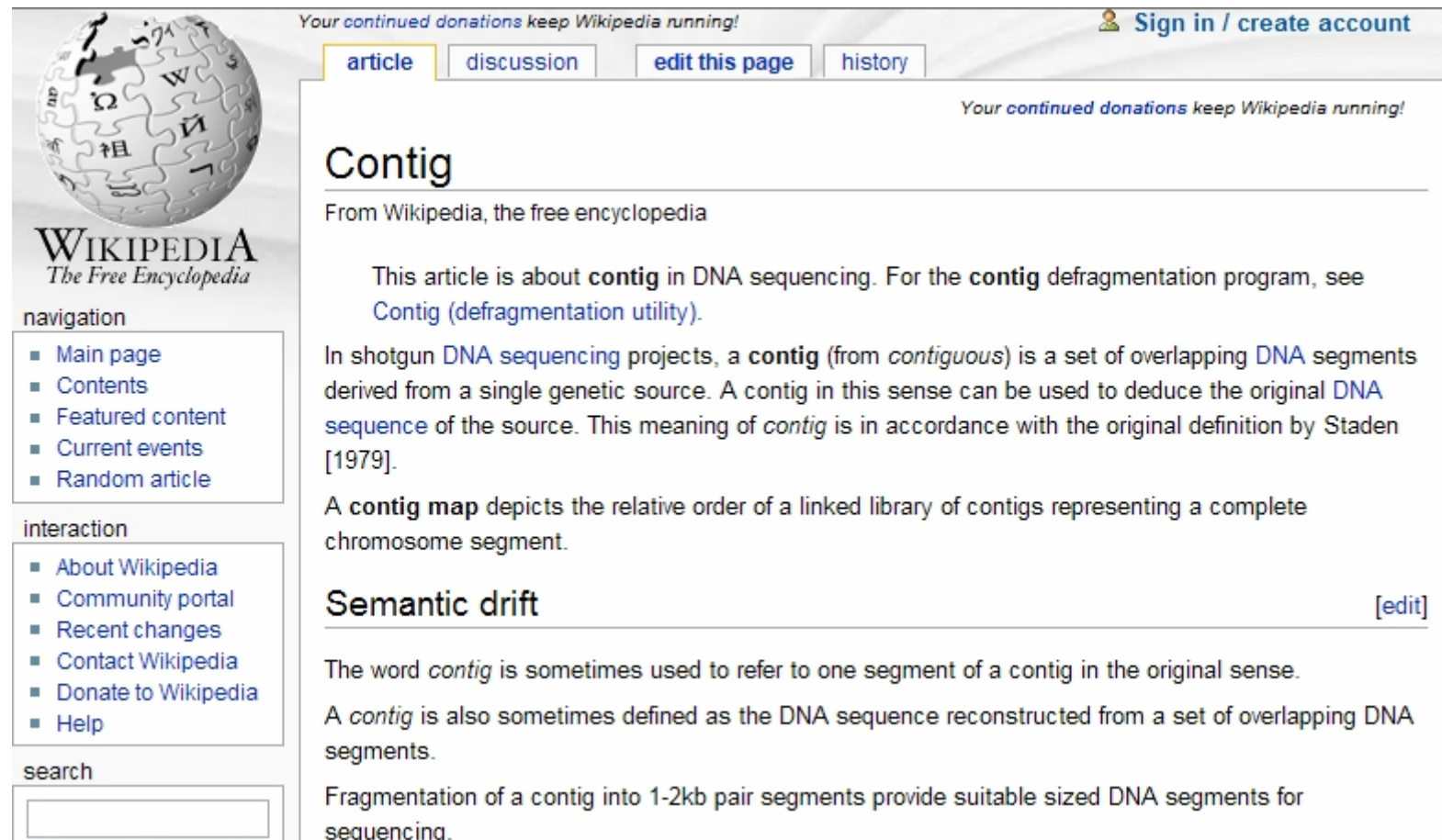
1'000'000  
slides like this one

# The basics

TCCTCGGGGAGCAAGGTACTCTACACATTCATAGCGGAGTTAAGATGGGTTTCCAAATTTT  
TCCAGCAAGGAAACGGCTCGGAGTCTTGAACACTGCTACCAATACAGCAACAGCTTA  
CTGGAACATAAAATCTCTGTATTTCAAATAACAGCCGCCCCACTACCACTAAGTGAAGT  
ATCCACCAACACACACAGCCACCACTTAAGCTTTTGTAGATCTCGGCTCGCTTTGGGGAACA  
GGTCTTTGAGAGAACTCTCTTTTAAGTGCAGAACAAAGGATTTTCATAGGCTCCAGCGTCG  
GTGCCCGAGGGCGCCCACTCCAAATGAGCTGGAGCAAAAAGGAAGGATGGGGGACTCTG  
GAGTAGGCATAGGGCGCGCCCTTCCAGACAGGCTGCTCGGACTCTTAAAGGTACAGCA  
GAAGAGAACAACACATCGGCTCTCCGCTTTATTTCGGTCAGATATGACGGTTGGGATGT  
TGACAGGAATTTCTCTTCTGCGCATGAGAAATACCCGACGGCCACCCAGGCTCTGCT  
TCTCGGGGTGGTGGCTGTGCTGAGCTGTGCGAGTCACGGCTGAGCGCCAGCGGGCTGT  
GTGGCGCGAGCTTTCTGAATCTAGGCGGACGGGCGGAGCCGCTCTGAGCACTCTGTGCGCT  
CTGCTGCGCCTCGGGTGGCTTTTTCGCGGGTGGTGGTGGCGCGGGGAAGCTGAGGGGA  
CAGATTTTGACCGCGCGGTTTTGTGAGCTTACTCGGCCAAAAAGAACTCGAGGCA  
TTGGACGGGTGTGGGGTGGTGGTGGTGGGACGAGCGCTCTTCGCAAGCTCCAGT  
CCAGCGGTGGTGGGGAGCGCTCACGCGGCTGCTGGGCGGCTCTGCGCTCTTGGCCTTTGT  
TCTTCTGCCAACCCCACTGCTCTGAGAGAAAGTCTTTCGCCAAGGAGATTTTCG  
CAAGCAAAATTCGAGCGCCGCCCTCTCCCTGGGTCTCCATTTCCGCTCTCGGCGCGGCT  
TTGGGCTCGGCTCTCAGCTCAAGACTTAACCTTCCGCTCCAGCTGCTCCAGATAGCGCAT  
CTGAAATTTCTTGAAACACGATACCTTTTAAACGAATATGCTGTTTGGGGAATTCGT  
TACAGCTCTTGGGCACGCTGTATTTGGCTTACTTAAGCCCTCGGTGAATTTGTGTATTTCG  
AAGACATGCTGATGGGAATTACAGGCGGCGTTGGTCTCTAACTGGAGCCCTCTGTCCCC  
ATAGGCCAGCGCTCACTGGTTAGCGTGTATTGAAACATAATCGTATGAAATATCTTCTCT  
TAGTGCACATAGCCACGTCTTAGTGCTTAATTTGGCTAGTGAGCAACGGTTTGACAGCT  
TAGTGTGTAATTAATCTCTCATPCTCCAGCTAAGCTGTACCCTTCAGGAGATGGGACTGA  
ATTAGAATTCAAACAAATTTTCCAGCGCTCTGTAGTTTAACTCCAGTACATATAAGA  
ATGCATCCCTGTGTAAGTGCAATTTTGGTCTCTGTTTTCGAGCTATTTTACCAAGCAT  
GGAGAAATCTCGTAGTTAAAAATGCTTATGGATCTCAAAGAGGCCAACATTTTGTAA  
ATTTTAAAGCAGCTGCAACAGGAGGATTGACAAATTTATATACCTTTATAAAT  
ACACGGAAGAAAGTTCTTAAAAAATGTCTGTCTAAAAACCCAGTAGCTCAATATGCT  
TTGAGCAATAAACTGTCTCTTATGTGTATGATAAATCAGTTAACACATAATCATCTGT  
TCGAGGTTTAAACCATGATAAATATAGAAGCTCTAGTGGAATAAGAGGAACTGGCCCT  
GAGTAGCTAGGAGAACTATACTAAACAACTAGAGCAATTAATGTACTTTATGAGCA  
TGATGTGCAACTTTTGGTTTCAGTACTCTTATPACTTTAAATAAGTCTAGGACCCCC  
GGAGTGTCTTTTGTATGTAGCTATACCATATGAATAATTAACATAGAATTTTAAGCT  
GGGCTGGTGGTCTACGCCCTGTAATCCGACCTTTGGGAGCCGAGGTGGCGGCTAC  
TTGAGCGCAGAAGTTTGAGACAGCGCTTGGCCAACTGGTGAACAGCTATCTCTACTAAAA  
ATGACAAAAAATGTCGCTGCTGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGT  
GAGGACAGAGATGCTGTATGTAACCTCGAGGACAGAGTTGCACTGAGTCAAGATGATGCCA  
CTGCACCTAGCTTGGCGCACATAGCATGACTCTCTCTCAAAACAACAAACAACAAAA  
AACAAGAAATTTAAAGTTAATTTACTTTAAAAATAAGAAAGTAAACCAATTCGATATAT  
CAACAACATCTTAGGAAAAATAACTTTTGAACAAAGAGTAGTGGAAATGTTTACATTT  
TTTGCACTCTCTCTTTAATGTCTGGCTAAATAGAGATAGCTGGATTTCTATCTGTGCTCT  
AATCTGTTATTTGGTGAAGAATGATGAAAAAAATTAACCTCAGCTTGAAAAAAGGAAT  
ATTTTAAATAGTTTTCAGTTACTTTTGGTATTTTCTCTGTACTTTTGACATAGATTTTCA  
AGAGTCTAATAGATGATTCACATAGGTTCTTCCCAATGTGCACCACTGAGTGAATTTATTT  
GGAAGATAGTAGGTGTTCTGAAATATACAAAGTTTCCAAATTTGTAATAATTCGATTAATTT

1'000'000  
slides like this one

# Some basic terms: Contig



The screenshot shows the Wikipedia page for 'Contig'. At the top, there's a navigation bar with links for 'article', 'discussion', 'edit this page', and 'history'. Below this, the article title 'Contig' is displayed, followed by the text 'From Wikipedia, the free encyclopedia'. The main content of the article explains that a contig is a set of overlapping DNA segments derived from a single genetic source, used to deduce the original DNA sequence. It also mentions that a contig map depicts the relative order of a linked library of contigs representing a complete chromosome segment. A section titled 'Semantic drift' explains that the word 'contig' is sometimes used to refer to one segment of a contig in the original sense, and that a contig is also sometimes defined as the DNA sequence reconstructed from a set of overlapping DNA segments. The page also includes a sidebar with navigation links and a search box.

Your continued donations keep Wikipedia running!

[Sign in / create account](#)

[article](#) [discussion](#) [edit this page](#) [history](#)

Your continued donations keep Wikipedia running!

## Contig

From Wikipedia, the free encyclopedia

This article is about **contig** in DNA sequencing. For the **contig** defragmentation program, see [Contig \(defragmentation utility\)](#).

In shotgun [DNA sequencing](#) projects, a **contig** (from *contiguous*) is a set of overlapping [DNA](#) segments derived from a single genetic source. A contig in this sense can be used to deduce the original [DNA sequence](#) of the source. This meaning of *contig* is in accordance with the original definition by Staden [1979].

A **contig map** depicts the relative order of a linked library of contigs representing a complete chromosome segment.

### Semantic drift

[\[edit\]](#)

The word *contig* is sometimes used to refer to one segment of a contig in the original sense.

A *contig* is also sometimes defined as the DNA sequence reconstructed from a set of overlapping DNA segments.

Fragmentation of a contig into 1-2kb pair segments provide suitable sized DNA segments for sequencing.

**navigation**

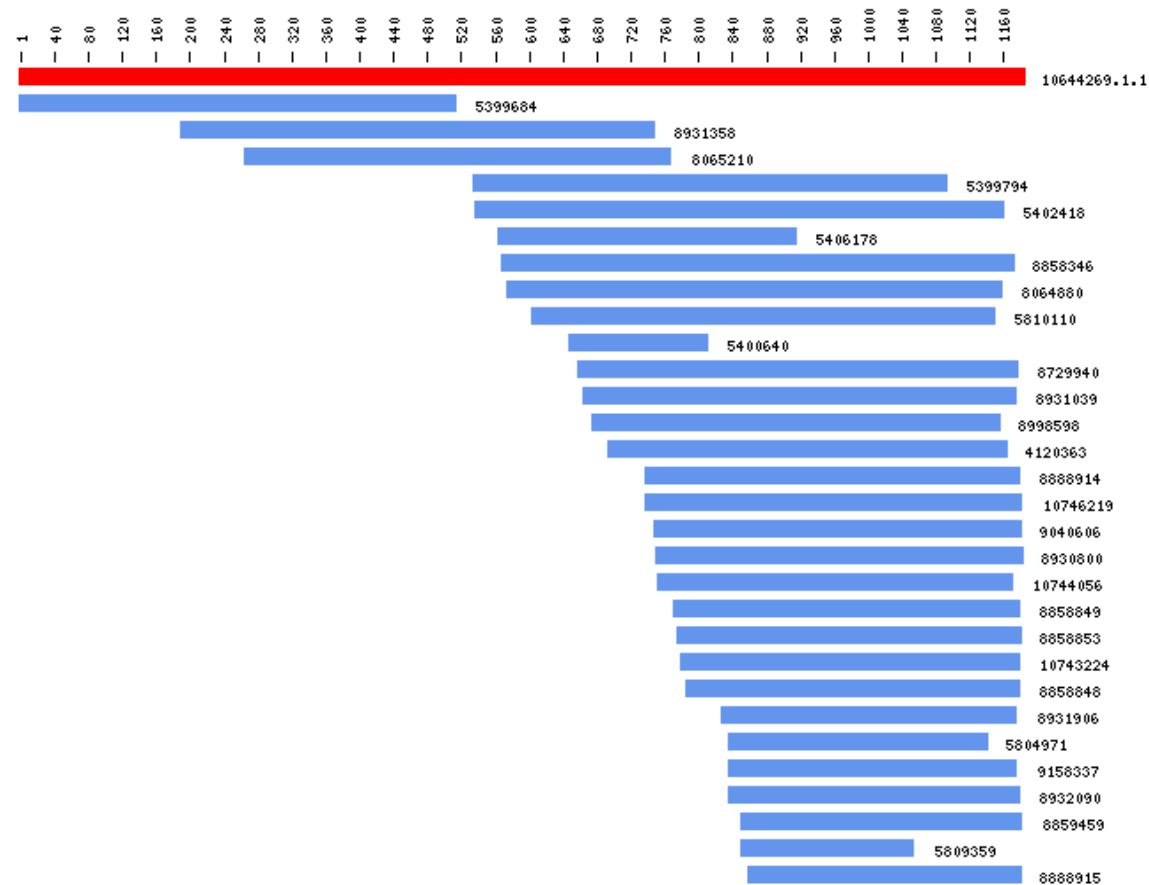
- [Main page](#)
- [Contents](#)
- [Featured content](#)
- [Current events](#)
- [Random article](#)

**interaction**

- [About Wikipedia](#)
- [Community portal](#)
- [Recent changes](#)
- [Contact Wikipedia](#)
- [Donate to Wikipedia](#)
- [Help](#)

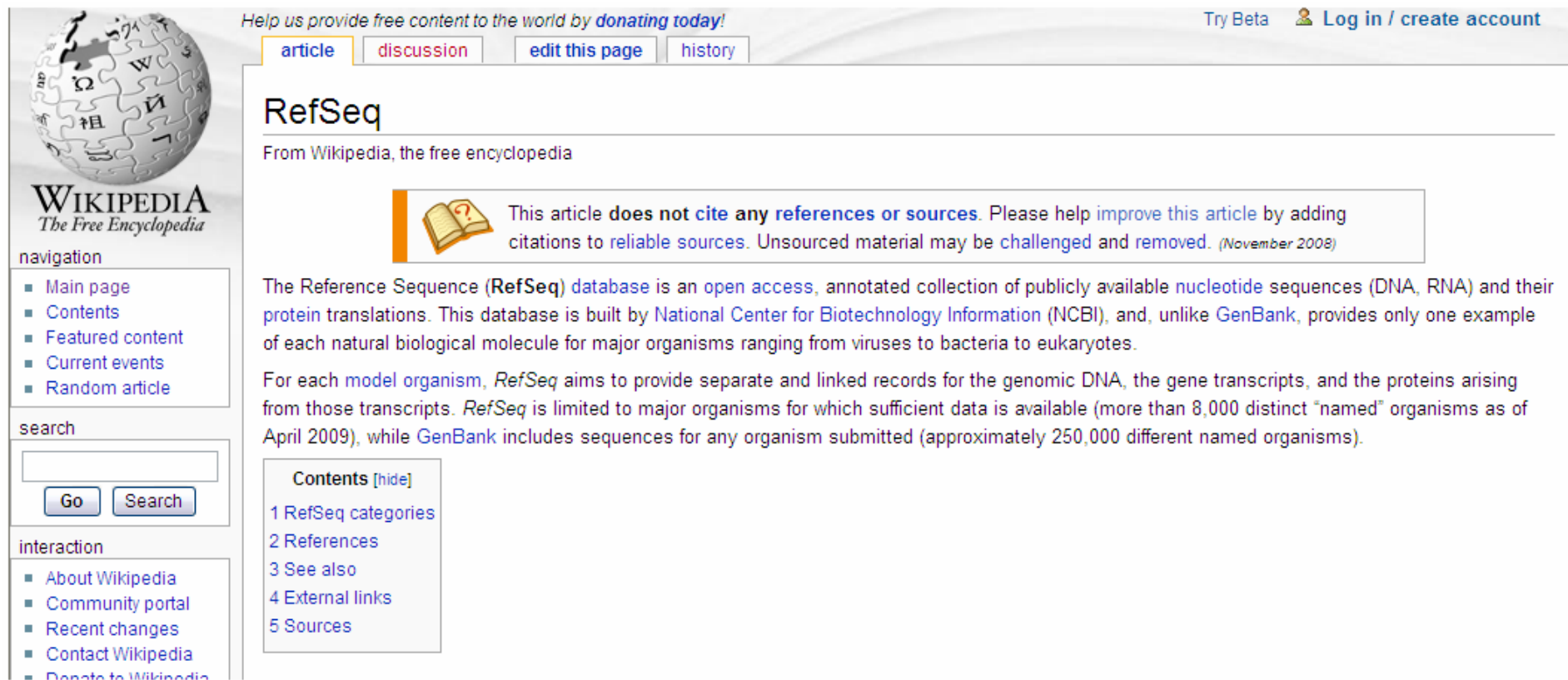
**search**

# Some basic terms: Contig





# Some basic terms: RefSeq




The screenshot shows the Wikipedia page for "RefSeq". At the top, there is a navigation bar with links for "article", "discussion", "edit this page", and "history". A banner at the top right encourages donations and offers a "Try Beta" option along with "Log in / create account" links. The article title "RefSeq" is prominently displayed, followed by the subtitle "From Wikipedia, the free encyclopedia". A warning box indicates that the article does not cite any references or sources. The main text describes the RefSeq database as an open access, annotated collection of publicly available nucleotide sequences (DNA, RNA) and their protein translations, built by the National Center for Biotechnology Information (NCBI). It notes that unlike GenBank, RefSeq provides only one example of each natural biological molecule for major organisms ranging from viruses to bacteria to eukaryotes. For each model organism, RefSeq aims to provide separate and linked records for the genomic DNA, the gene transcripts, and the proteins arising from those transcripts. RefSeq is limited to major organisms for which sufficient data is available (more than 8,000 distinct "named" organisms as of April 2009), while GenBank includes sequences for any organism submitted (approximately 250,000 different named organisms). A "Contents" sidebar on the right lists sections: 1 RefSeq categories, 2 References, 3 See also, 4 External links, and 5 Sources. The left sidebar contains navigation links such as "Main page", "Contents", "Featured content", "Current events", "Random article", and a search bar.

Help us provide free content to the world by [donating today!](#) Try Beta [Log in / create account](#)

[article](#) [discussion](#) [edit this page](#) [history](#)

## RefSeq

From Wikipedia, the free encyclopedia

 This article **does not cite any references or sources**. Please help improve this article by adding citations to [reliable sources](#). Unsourced material may be challenged and removed. *(November 2008)*

The Reference Sequence (**RefSeq**) [database](#) is an [open access](#), annotated collection of publicly available [nucleotide](#) sequences (DNA, RNA) and their [protein](#) translations. This database is built by [National Center for Biotechnology Information](#) (NCBI), and, unlike [GenBank](#), provides only one example of each natural biological molecule for major organisms ranging from viruses to bacteria to eukaryotes.

For each [model organism](#), *RefSeq* aims to provide separate and linked records for the genomic DNA, the gene transcripts, and the proteins arising from those transcripts. *RefSeq* is limited to major organisms for which sufficient data is available (more than 8,000 distinct "named" organisms as of April 2009), while [GenBank](#) includes sequences for any organism submitted (approximately 250,000 different named organisms).

**Contents** [\[hide\]](#)

- 1 RefSeq categories
- 2 References
- 3 See also
- 4 External links
- 5 Sources

**navigation**

- [Main page](#)
- [Contents](#)
- [Featured content](#)
- [Current events](#)
- [Random article](#)

**search**

[Go](#) [Search](#)

**interaction**

- [About Wikipedia](#)
- [Community portal](#)
- [Recent changes](#)
- [Contact Wikipedia](#)
- [Donate to Wikipedia](#)

# This is the RefSeq, based on contigs

CCTCCGCGGACGAGGTACTCTACACTTCTAGCGAGGTATGATGAGGGTTTACCAATATTTT  
 TCAACGAGGAAGAACCTGCTGGAGGTCTTGAAACCTCTGCTCCCAATAGCAACAGCTTA  
 CTGGAACATAAAATCCTCTGATTTCAAATAACAGCCGCCGCCACTACCACTAAGTGAAGTC  
 ATCCACCAACACACACGCCACCACTCTAAGCTTTTGTAGATAGCTGCTCTTGGGGTGAAC  
 GTCTCTTGAGAGAACATCCCTTTTAAAGTGCAGAACAAAGCTTTTCATAGTCCAGGACGTG  
 GTGCCCGAGGGCGCCCAACCAATACAGCTGGAGCAAAAAGAAAGGGATGGGGGACTTGT  
 GAGTAGGCACTAGGGGCGGCCCTTCAAGCAGGTGGCTGGGACTCTTAAAGGTCACGGA  
 GAAGAAGAACCACTACGAGCTCCGGCTTTATTTCGGTCAGATAGCTGAGGTGGGAATGT  
 TGACAACGAATTTCTCTTCCGACACTGAGAAATACCCGACGGCCGACCCAGCTGGGTGAC  
 TCTCCGGGTGTGGCTGTGCTGCGCTGCGCTCGCGCTACGGCGTACGTGACGCGCGGCTT  
 GTGCGCGGAGCTTCTGAAATAGGCGCGACGAGCGGAGCGCTGTGTCACCTGCTGCGCTCT  
 CTGCTCGGCTCCGGTGCTTTTTCGCGCGGTGGTTCGCGCGGGGAGAAGCTGAGGGGA  
 CAGATTTGTGACCGCGCGGTTTTTGTACGCTTCTCTCGGCAAAAAGAACTGCACCTCT  
 TGGAGCGGGTTAGTGGTGGTGTAGTGGTTGGGACGAGCGCTCTTCGACAGTCCCAAGT  
 CCACGCTGCGGGGAGAGCGCTACGCGCGGGGCGCTGCTGCGCGCTTCTTGCCCTTTTG  
 TCTCTGCCAACCCCACTGCTGAGAGAAAGTCTTGTCCCGAAGGCAGATTTTTCG  
 CAAGCAAAATCGAGGCCCGGCTCTCCCTGGGTCTCAGATTCCCGCTCCGGCGCGCGCT  
 TTGGGCTCCGCTTCAGCTCAAGCTTAACCTCCCTCCAGCTGTCCAGATGACGCGCAT  
 CTGAAATTTCTTGGAAGACGATCACTTAACTGGAATTTGTGTATTGGTGAAGTGTCTT  
 TACAGATCTCTGGCAGCGTGTATTTGCTCTTATAGCCCTTGTGATTTAGGCTGTATTCCG  
 AAGACATGCTGATGGAATTAACAGCGCGCGTGTGGTCTTAACTGGAGCCCTCTGTGCCCC  
 ACTAGGCGCGGCTCACTGGTTAGCGTGTAATGAAACATAAATCTAGAAATCTGCTACCA  
 TAGTGCACATGGACAGTTCTTCAGTGTGATTTAGTTGGTATGTGGCAGCGGTTTGACGACGA  
 CAGCTGTAAATTTCTCCATCTACAGTAACTGCTTACCGTTCCAGGAGATGGGACATCT  
 ATTAGAATTCAAAAGTAATTTTCAGCGCTTCTAGTTTAACTCTCAGTCAACATAAAGAA  
 ATGCACTCCTGTGTAGTGCAATTTTGGTCTTCTGTTTTCAGACATTAATTACCAAGCATTT  
 GGAGGAATATCTGATGAGTAAAAATGCTATTTGATCAACAAAGAGGCCAACATTTTGTAA  
 ATTTTAAAGCAGCTGCAACAAAGCAGGTATGCAAAATTTTATATCTATTATAATTT  
 ACACCGAGAAAGTGTTTCTAAAAAATCTTGCTAAAAACCCAGTCAACATCACTGTCTT  
 TTGAACCACTAAATCTCTCTTATGTGTGTAATAATCCAGTTAAACACATAATCACTGTT  
 TGCAGGTAAACCATGATAAAATATAGACGCTTAGTGTAATAAGAGAAATCGGCCCTAC  
 TGACTAGCAGTAGGAACAATTAACATAACAACTAGAAGCATTAATGTATCTTATGGGACA  
 AGTTGTCCAACTTTTGGTTTTCAGTACTCTTAATCTTAAAAAGATCTAGGACCCCC  
 GGAGTGTCTTTTATGTAGCTTACCATTATAGAAATTTAAAACTAAGAATTTAAAGCT  
 GGGCGTGGTGAGCTACAGCTGTAACTCCGACACTCTGGGAGCGCGAGTGGGCGCATTAAC  
 TTGAGGCAGAGATTTTGAGCAGCTCCGCAACATGTGGTAAACCTTACTTCTACTAAAA  
 ATACAAAAAATGTGCTGCGTGGTGTGGTGTGGCTGCTGTAATCCAGCTACACGGGAGTGT  
 GAGGCAAGGAGAAATCTGTAACCTTGAGCGCAGGCTGTACGTGAGCAAGATCAATGCCA  
 CTGCACTTAGCTGGGCCACTAGCATGACTCTGCTCTCAAAACAAAAACAAACAAAA  
 AACTAAGAATTTAAGTAAATTAATCTTAAATAATTAAGAAGTAAACCAATTCATATATAT  
 CACAAAGCTTTTAGGAAAAATTTTTTTGAAGAACAGTAGTGGAATAGTTTTCATCTT  
 TTTGCAAGTCTCTTTAAATGCTGGCTAAATAAGAGATAGCTGGATCACTATCTGTGGTCT  
 AATCTGTATTTTGTGTAAGATATGTAAAAAAATTAACCTCACGTTCAAAAAGGAAT  
 ATTTTAATAGTTTTCAGTACTTTTGTGATTTTTCCTGTACTTGCATAGATTTTTCAT  
 AAGATCTAATAGATATACCATAGTCTTCCCAATGTGCAACATCACTGACGATGATTTTCA  
 GAAGATAGTGGTGTCTGAATTAACAAGATTTTCCAAATTTGTAATAATTCGATATATTC

# DNA variations

- Mutations
- Rare variants
- Polymorphisms
- Classically:
  - Silent changes (isocoding changes) [TGT > TG<sup>C</sup>, or Cys>Cys]
  - Missense changes [TGT > TG<sup>G</sup>, or Cys>Trp]
  - Nonsense changes [TGT > TG<sup>A</sup>, or Cys>End]
  - Others [for example: TGT > TG<sup>-</sup>, or Cys>fs]

# Mutations

- Rare DNA changes ( $<1\%$  allele frequency) that are associated to a (deleterious) phenotype

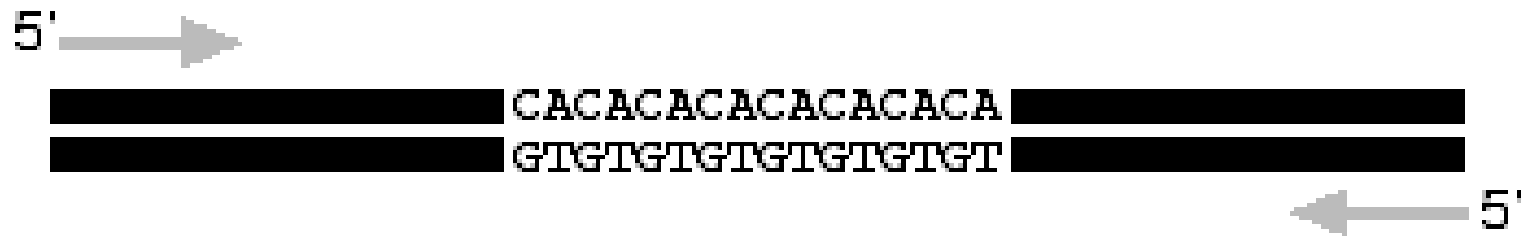
## Rare variants or rare changes

- Rare DNA changes ( $<1\%$  allele frequency) that are usually neutral

# Polymorphisms

- Common DNA changes ( $>1\%$  allele frequency)
- They are responsible for common phenotypic variations
- They include several kinds of DNA changes
  - Short tandem repeats (STRs or microsatellites)
  - Single-nucleotide polymorphisms (SNPs)
  - Polymorphic microdeletions or microinsertions
  - Copy number variations (CNVs)
  - Other changes

# Microsatellites



CACACA  
CACACACA

3,4



CA  
CACACACA

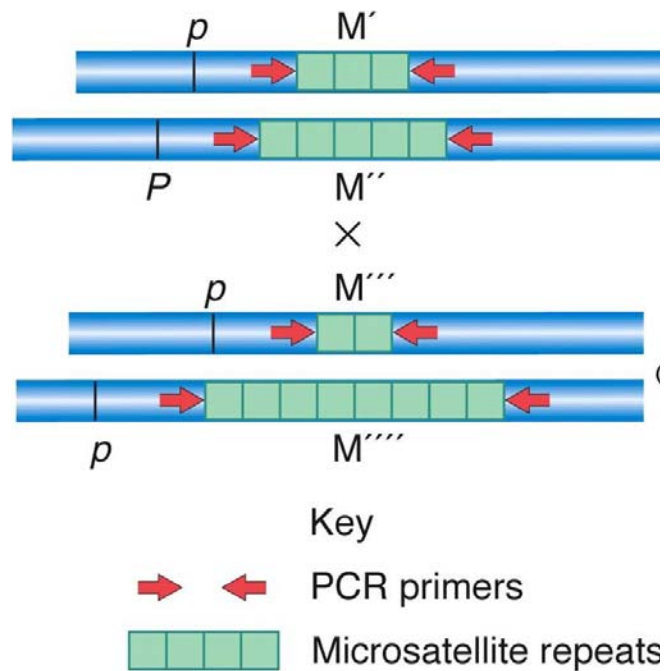
1,4



CA  
CA

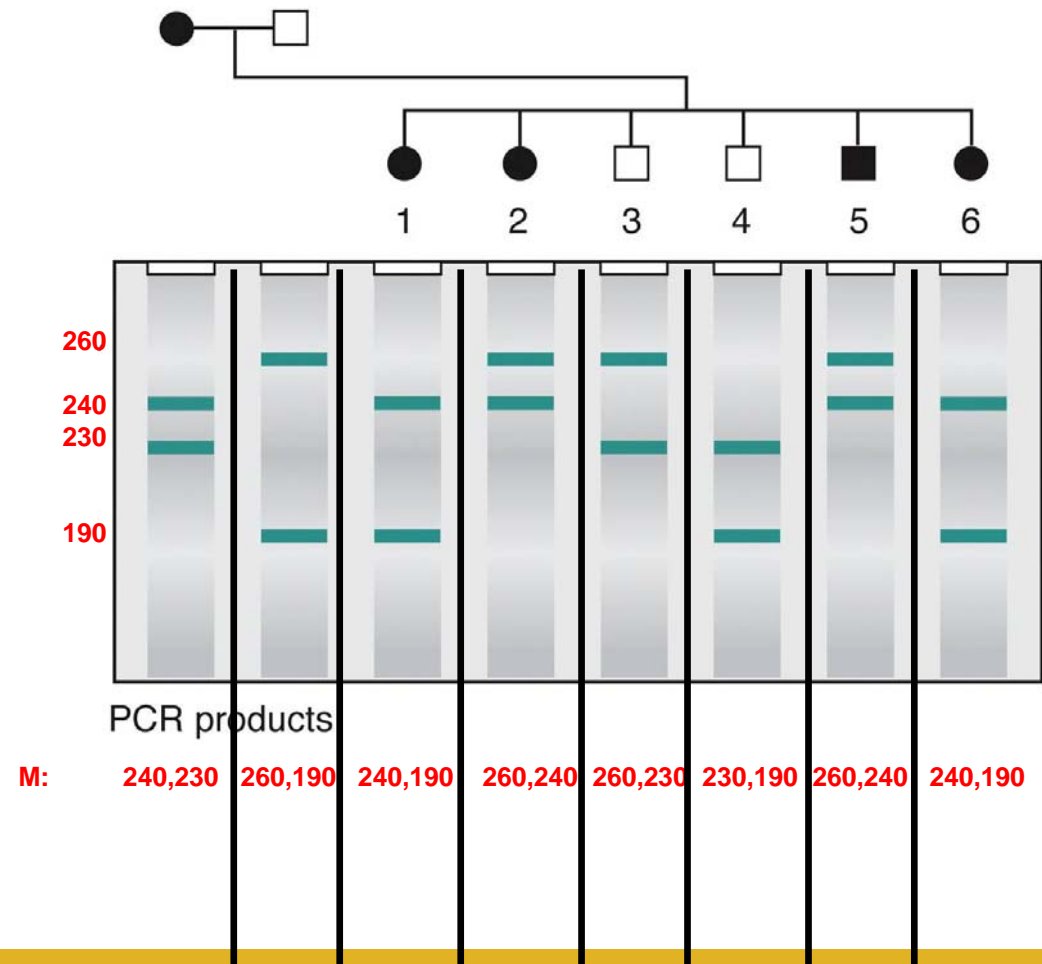
1,1

# Microsatellites



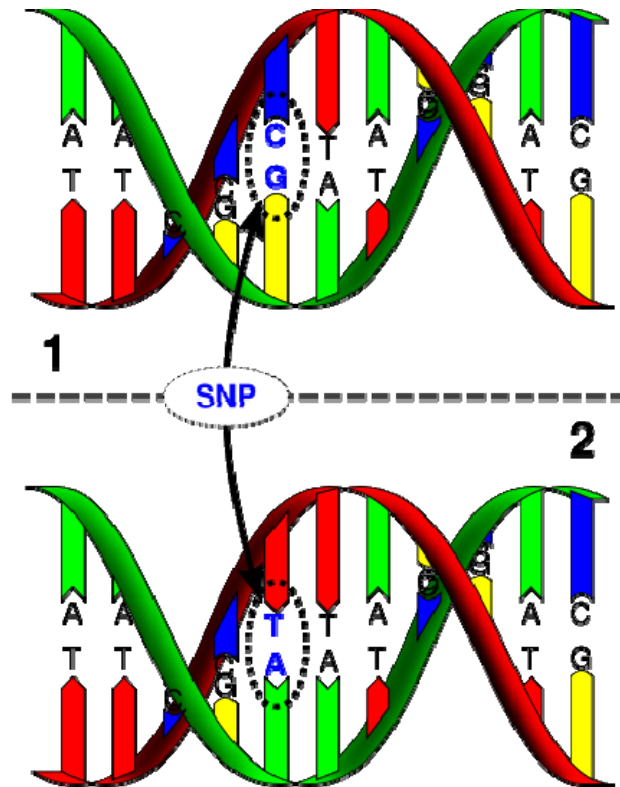
$M'$ ,  $M''$ ,  $M'''$ ,  $M''''$  = alleles of  $M$

© Memorial University of Newfoundland





# SNPs



AACG<sup>G</sup><sub>A</sub>ATCCAC

# Microdeletions or microinsertions

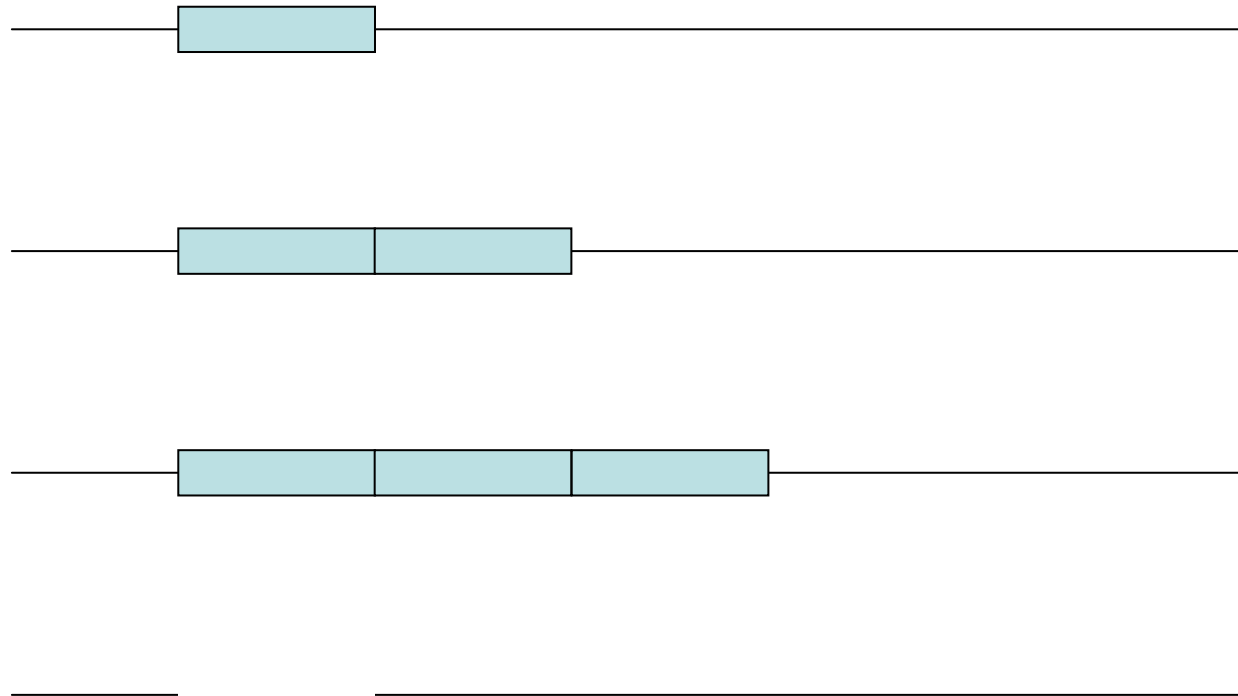
AGTGCTTGCAACTGACTAC

AGTGCTT--AACTGACTAC

AGTGCTTGCAACTGACTAC

AGTGCTTGC**C**AACTGACTAC

# CNVs



# DNA Markers

- DNA markers are sequences on the DNA that are variable and easily recognizable
- Usually, DNA markers are used as tools for genetic investigations
- In principle, any detectable polymorphism can be a DNA marker, but mostly **only SNPs and microsatellites are used.**

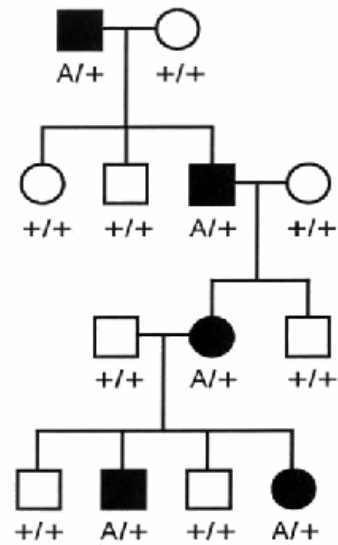
**Important note: from now on, only Mendelian characters are considered...**

...and human genetic diseases are taken as example.

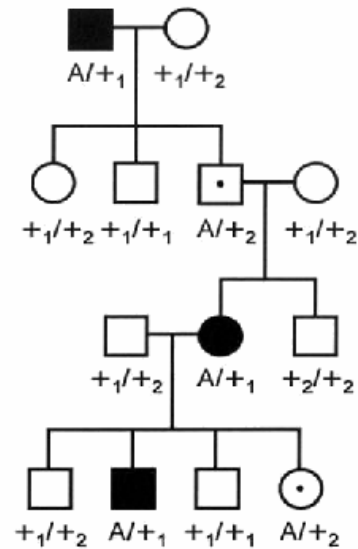
**HOWEVER**

the notions outlined here are in principle valid for other traits, other organisms, and other phenotypes

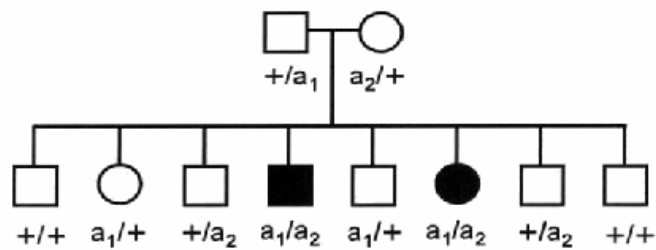
Autosomal Dominant



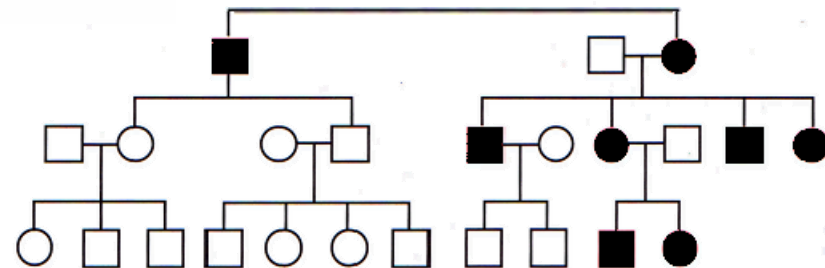
Autosomal Dominant (reduced penetrance)



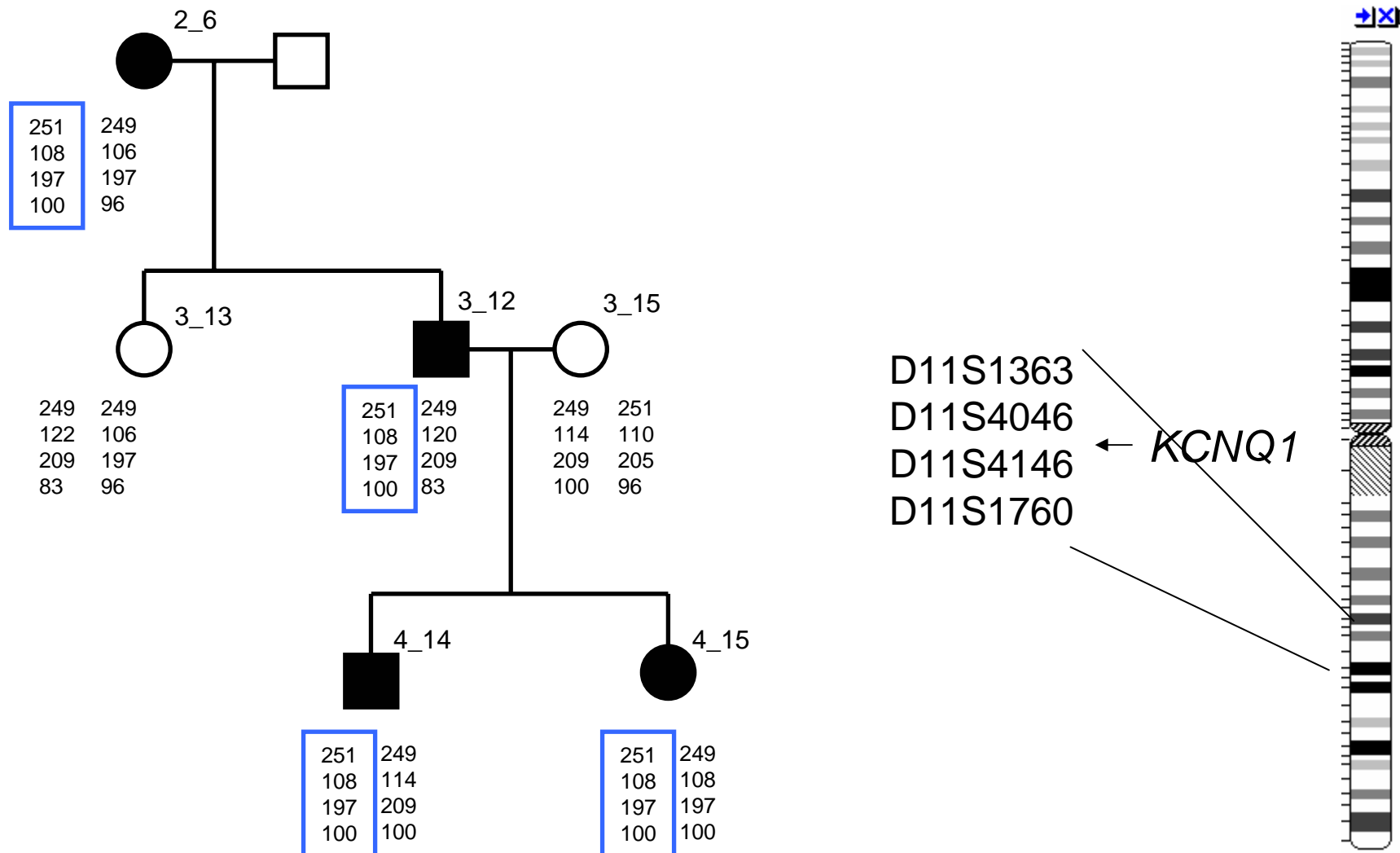
Autosomal Recessive



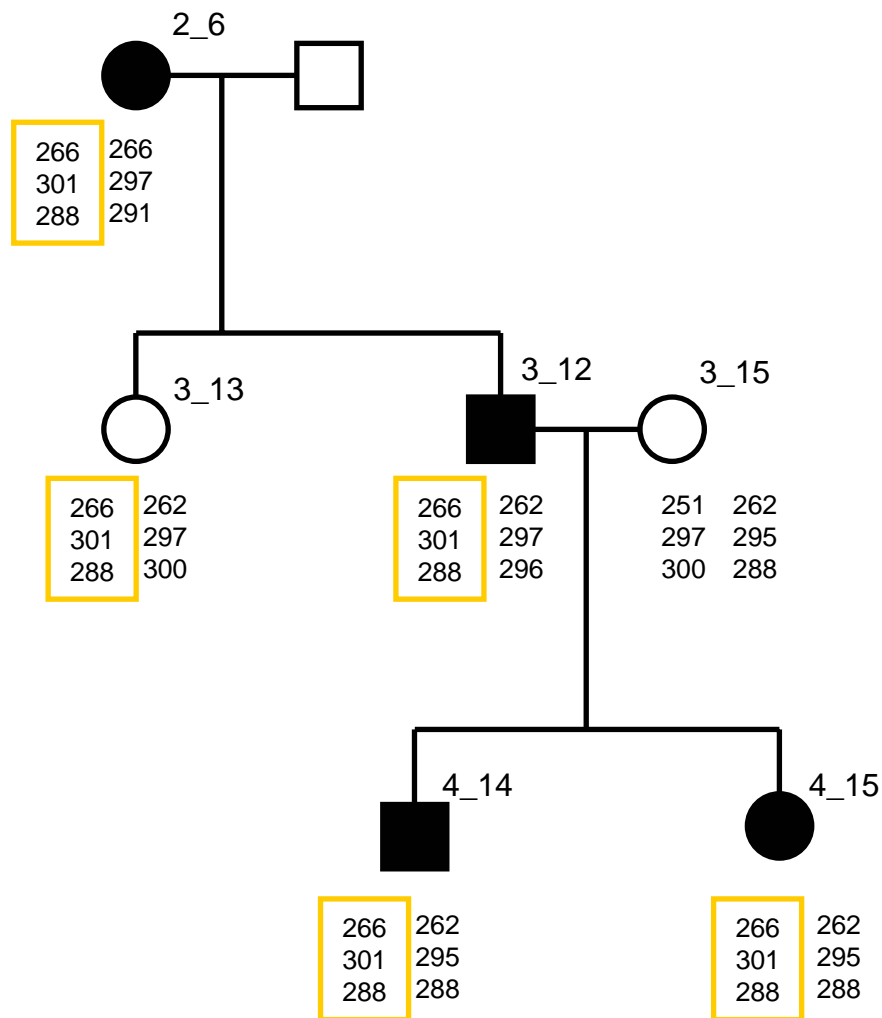
Mitochondrial (maternal)



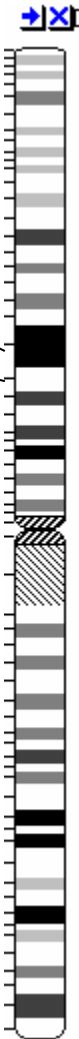
# Microsatellite mapping



# Microsatellite mapping



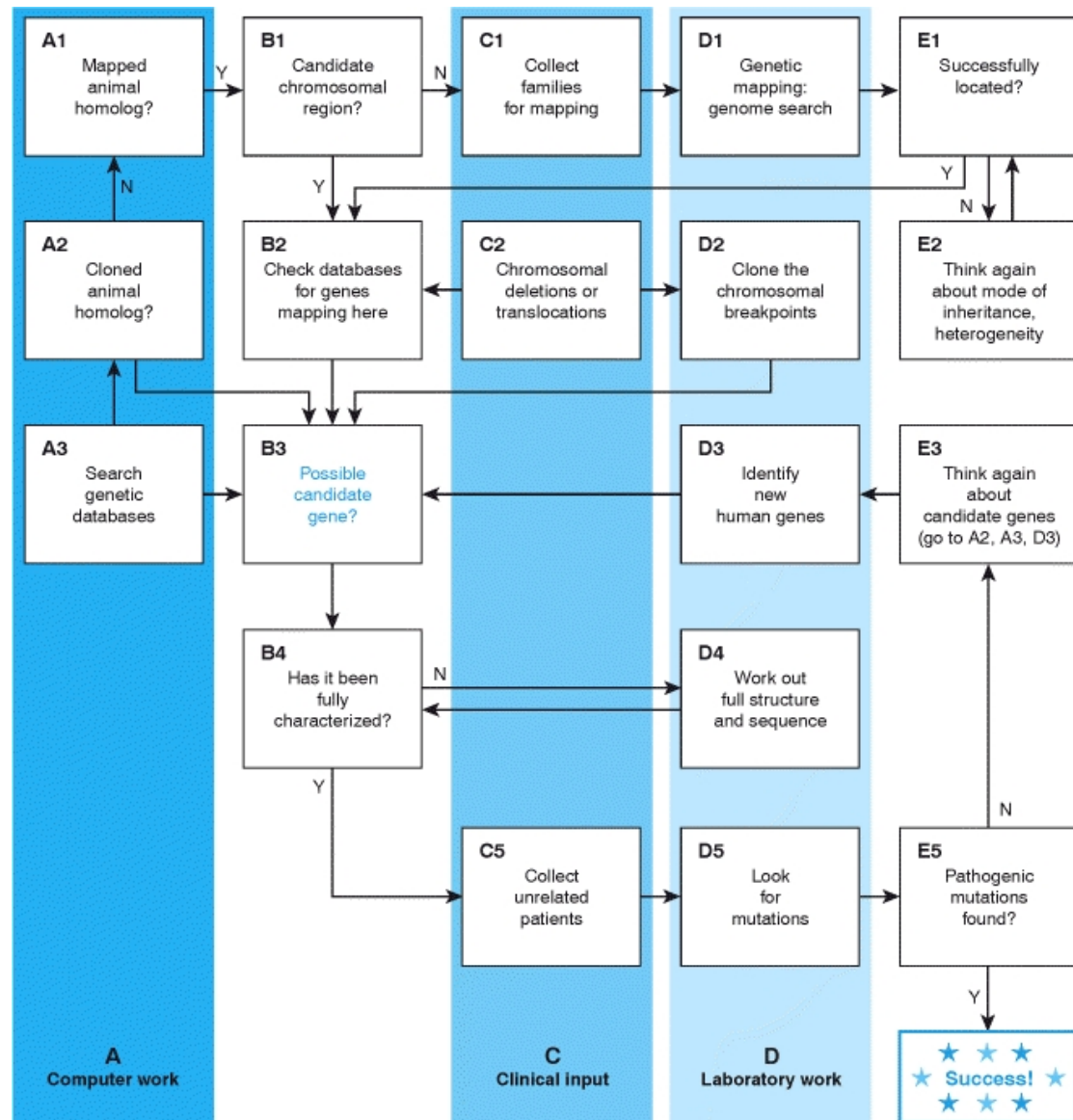
D3S1609  
D3S1277 ← *SCN5A*  
D3S3521





# Let's get (finally) started

- We got the sequence: now what?
- The identification of human disease-genes is taken as an example of experimental annotation of the raw DNA primary sequence
- A few cases are provided, to illustrate various real-life possibilities

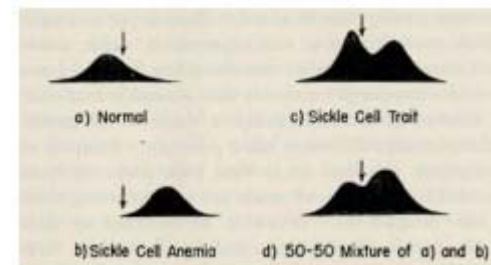


From: Strachan and Read  
Human Molecular Genetics

# Position-independent gene identification

## [1. Starting from the protein product]

- It is mostly a pre-genomic strategy, relying on protein information and on biochemical notions
- The most famous example is Sickle Cell Anemia, where hemoglobin was shown to be different in patients vs. controls (1949)

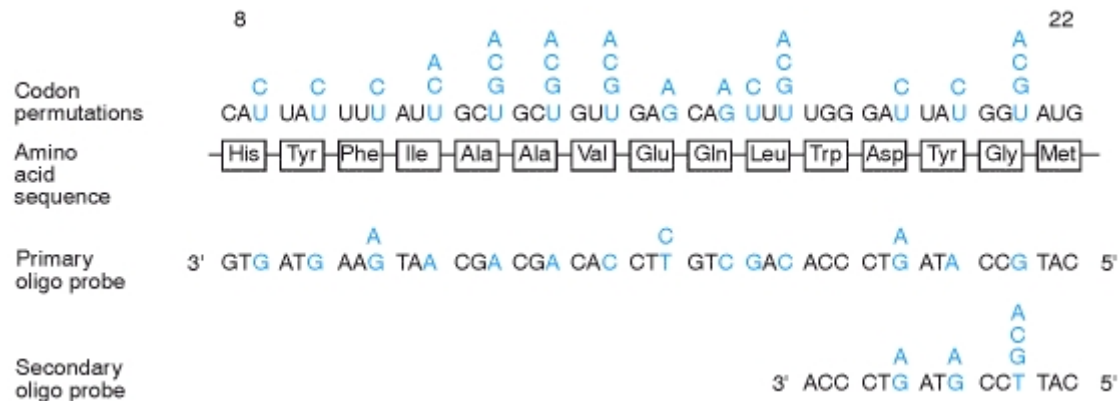


- Obviously, the genetic defect should lie in the DNA encoding for the globins

# Position-independent gene identification

## [1. Starting from the protein product]

- Classically, the identification of the gene, starting from the protein sequence, was obtained by reverse-translation of the aa sequence, followed by Southern blot



From: Strachan and Read  
Human Molecular Genetics

# Position-independent gene identification

## [1. Starting from the protein product]

- Nowadays, we would BLAST public databases with the sequence of interest...
- ...and the whole experimental strategy would last only a few minutes

# Position-independent gene identification

## [2. Starting from an animal model]

- Another (relatively rare) possibility comes from the identification of the gene in a mouse model and subsequent identification in another species (e.g. human)
- Again, in the past this was done with DNA probes. Today, we would again use BLAST.

# Position-independent gene identification

## [3. The “candidate gene” approach]

- The choice of candidates is based on:
  1. Tissue specificity and abundance
  2. Chromosomal location
  3. Sequence information and length
  4. Similarity to other known disease-causing genes
  5. Other characteristics (gene expression, etc.)

# Position-independent gene identification

## [3. The “candidate gene” approach]

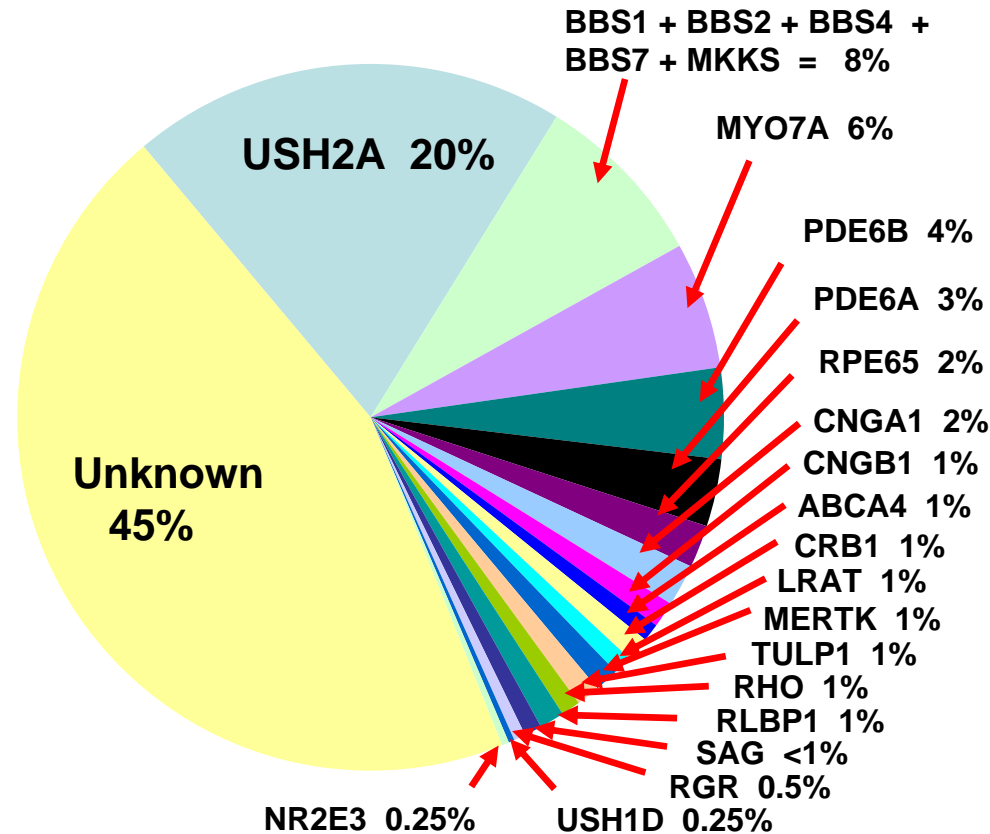
- The candidate gene approach represents a very powerful gene identification system. It is applied when all of the following conditions are fulfilled:
  1. The phenotype is genetically heterogeneous
  2. The molecular mechanisms determining the phenotype are known
  3. The sequence of the candidate gene is known
  4. Large cohorts of unrelated individuals are available



# Position-independent gene identification

## [3. The “candidate gene” approach]

Recessive retinitis pigmentosa genes (% of cases)

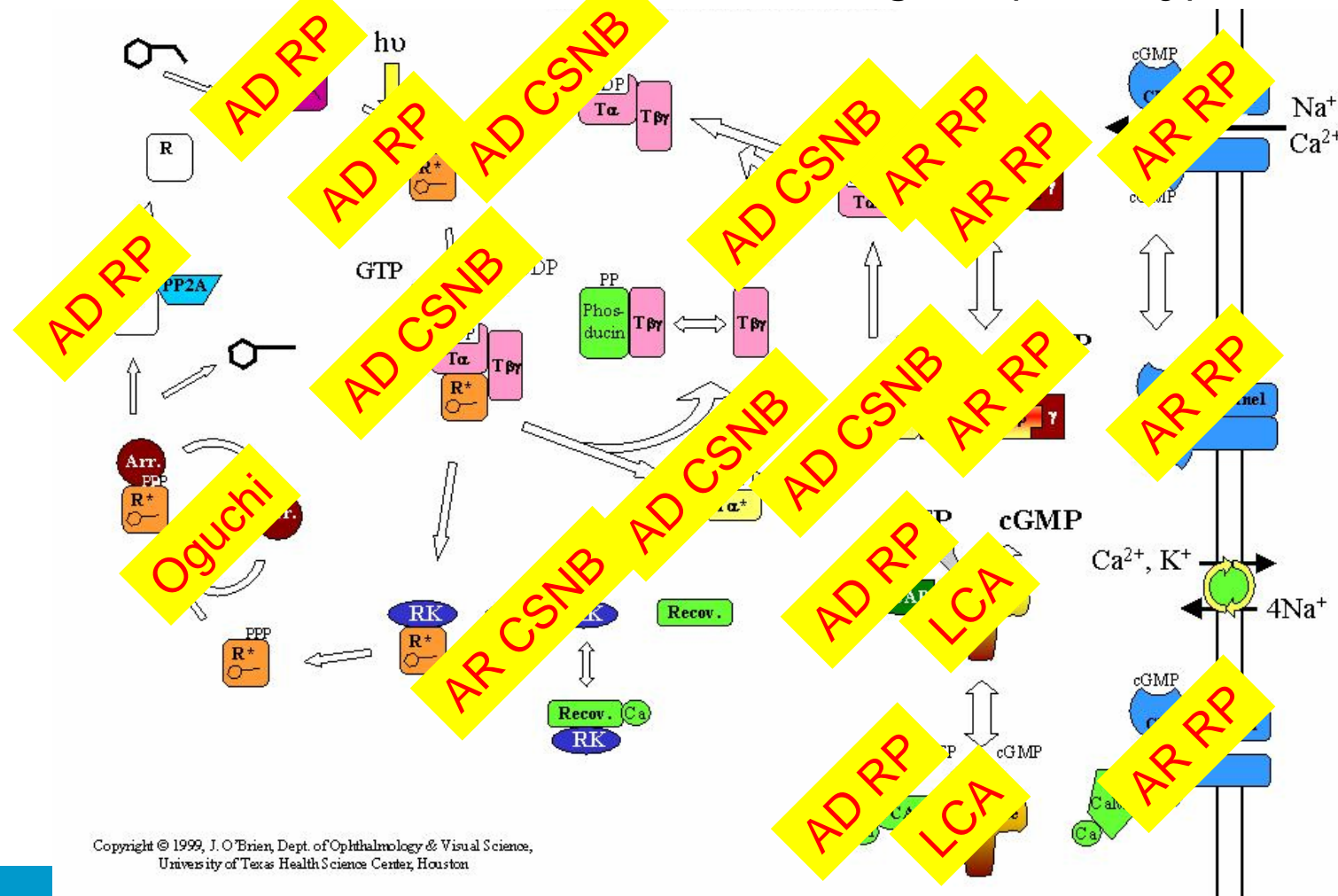


The phenotype is genetically heterogeneous

# Position-independent gene identification

## [3. The “candidate gene” approach]

The molecular mechanisms determining the phenotype are known

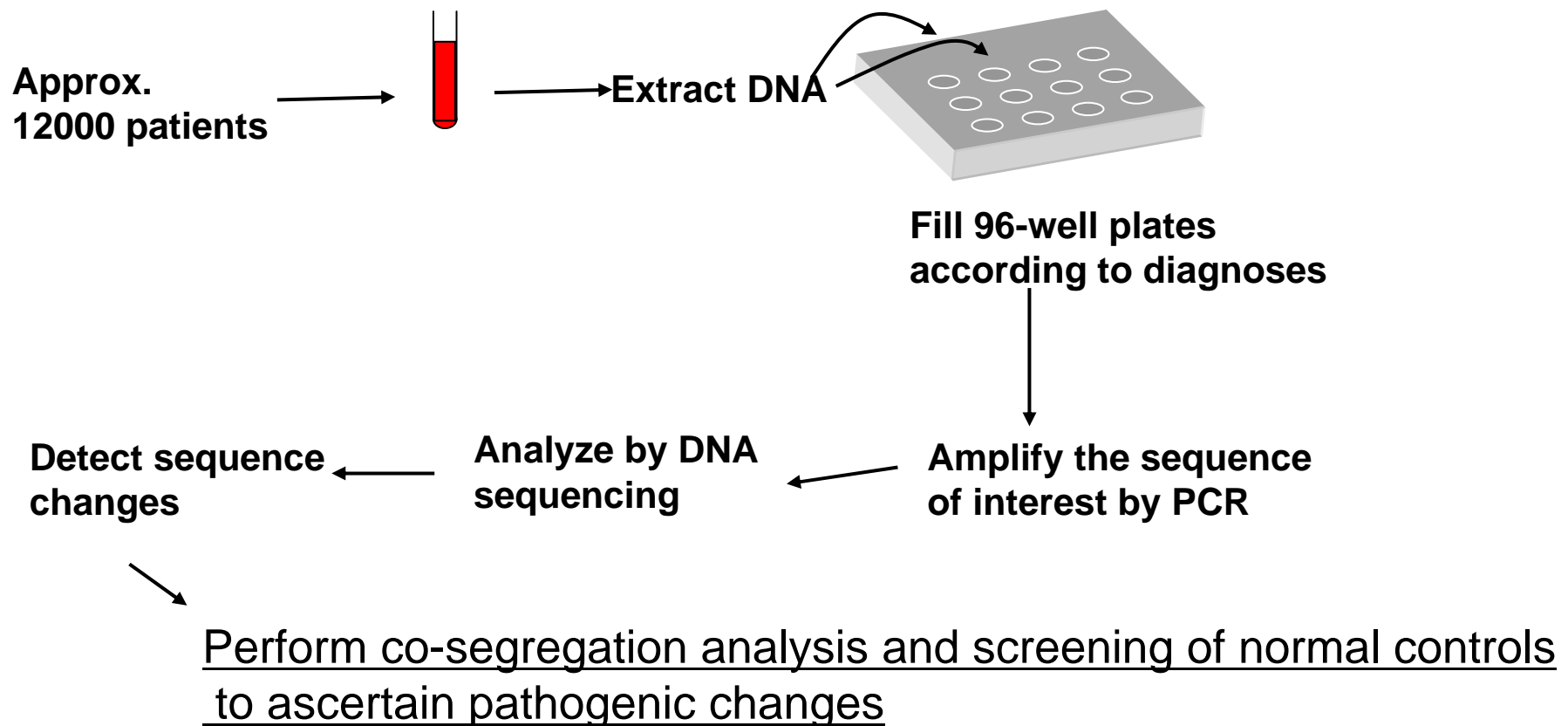


Copyright © 1999, J. O'Brien, Dept. of Ophthalmology & Visual Science,  
University of Texas Health Science Center, Houston

# Position-independent gene identification

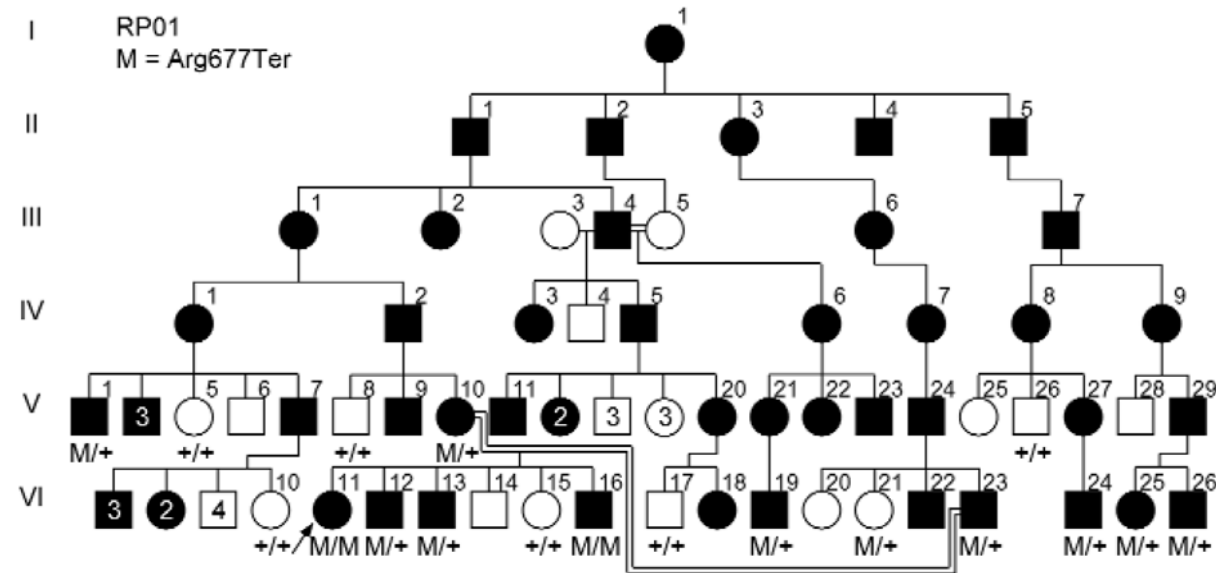
## [3. The “candidate gene” approach]

- The sequence of the candidate gene is known
- Large cohorts of unrelated individuals are available



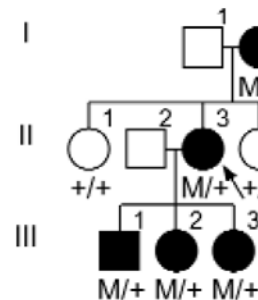
# Position-independent gene identification

## [3. The “candidate gene” approach]

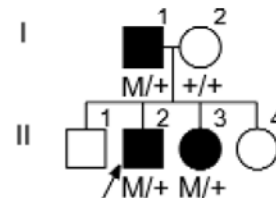


*b*

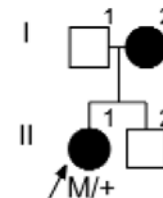
6886  
M = Arg763(4-bp del)



7050  
M = Leu762(5-bp del)



2480  
M = Leu762(5-bp del)



# Position-independent gene identification

## [3. The “candidate gene” approach]

	Have the DNA change	Do not have the DNA change	Total
Patients	5	91	96
Controls	0	200	200

Fisher's Exact Test p-value = 0.003

Chi-Square Test p-value = 0.005

# Position-independent gene identification

## [3. The “candidate gene” approach]

	Have the DNA change	Do not have the DNA change	Total
Patients	5	91	96
Controls	3	197	200

Fisher's Exact Test p-value = 0.117

Chi-Square Test p-value = 0.144

# Position-independent gene identification

## [3. The “candidate gene” approach]

- Further validation could include
  1. Biochemical analysis of recombinant protein products carrying the detected variant (e.g. phosphorylation assay for a wild-type vs. mutant kinase)
  2. Immunocytochemistry of the same (does the recombinant protein de-localize?)
  3. Construction of animal models
  4. ...

# Positional cloning (position-dependent gene identification)

## [Step 1: define the position]

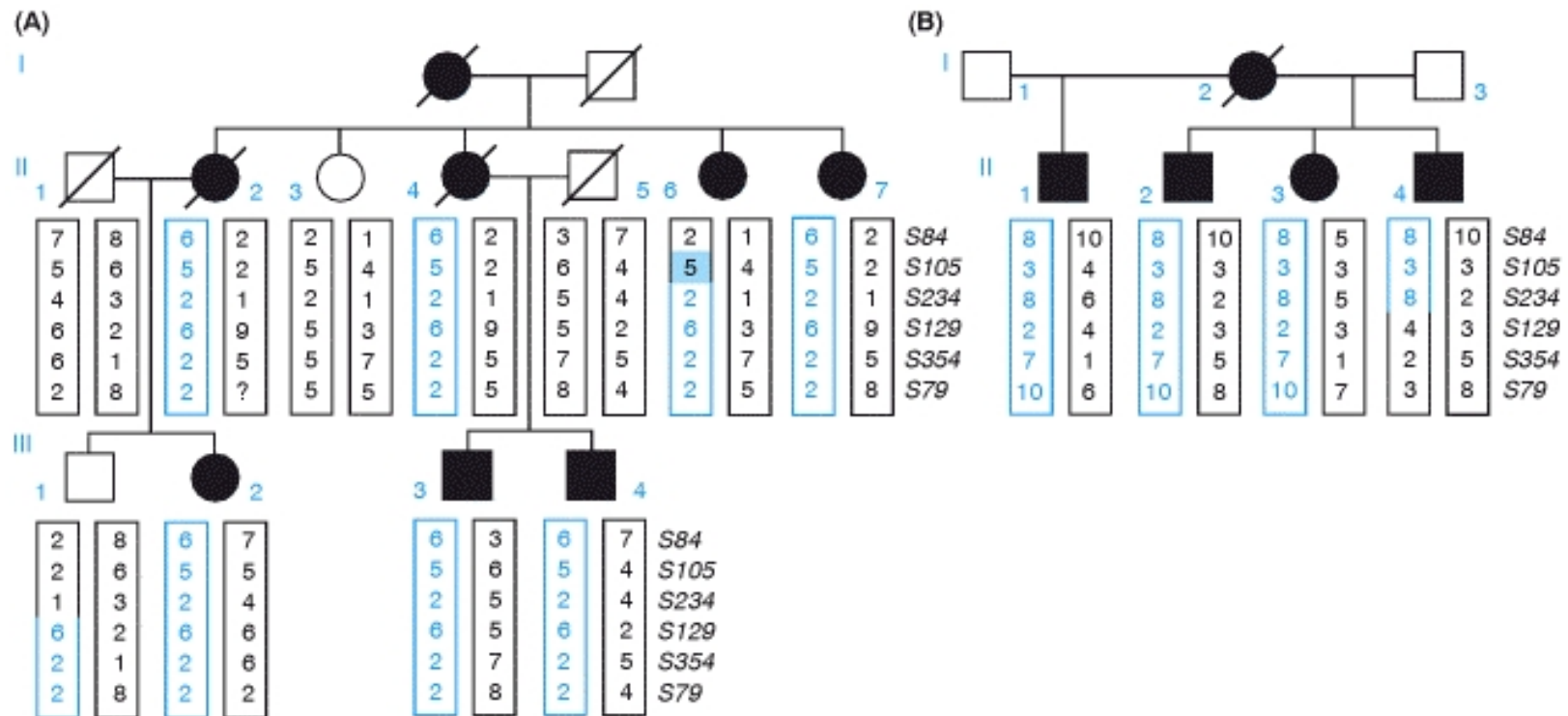
- Several methods are possible:
  1. Linkage and haplotype analyses (from families with multiple affected individuals). This is the most used method
  2. Synteny maps
  3. Chromosomal anomalies
  4. ...



# Positional cloning

## [Step 1: define the position]

## Linkage and haplotype analyses

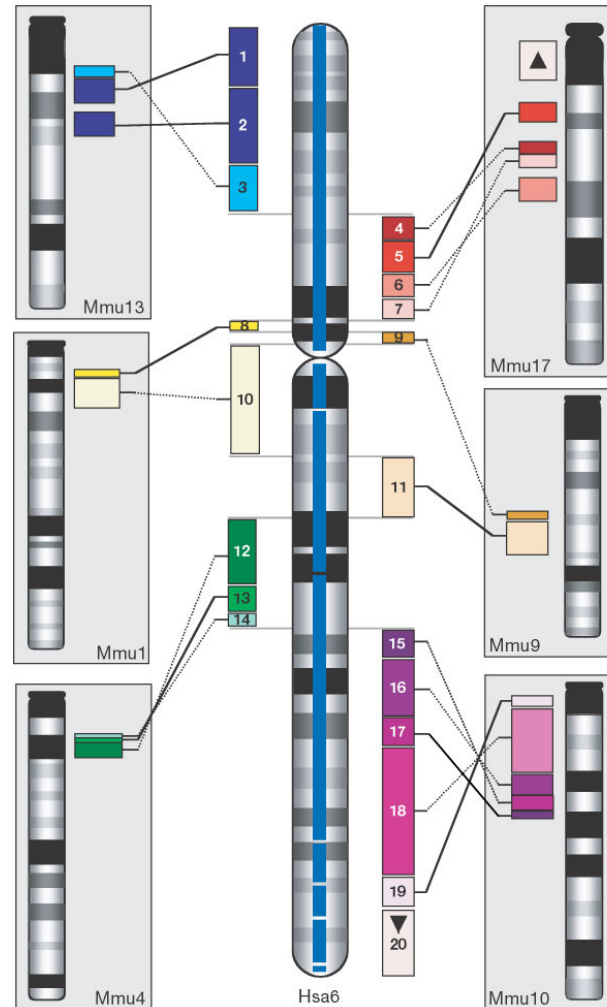


From: Strachan and Read  
Human Molecular Genetics

# Positional cloning

## [Step 1: define the position]

Synteny maps

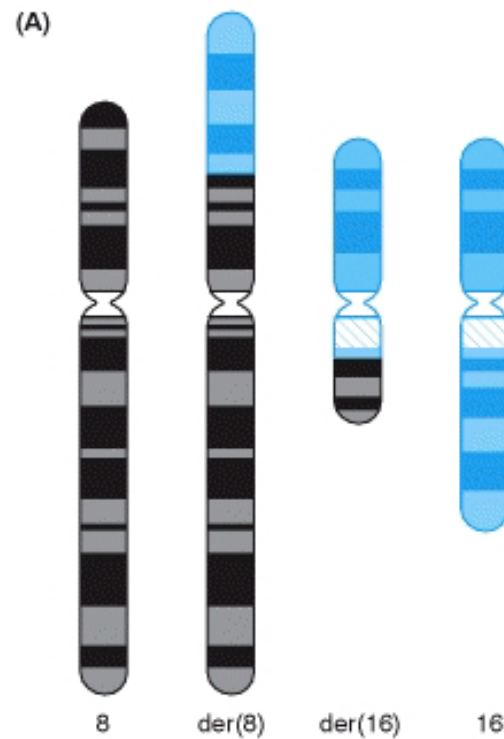


From: Gregory et al. Nature 418:743-750

# Positional cloning

## [Step 1: define the position]

### Chromosomal anomalies

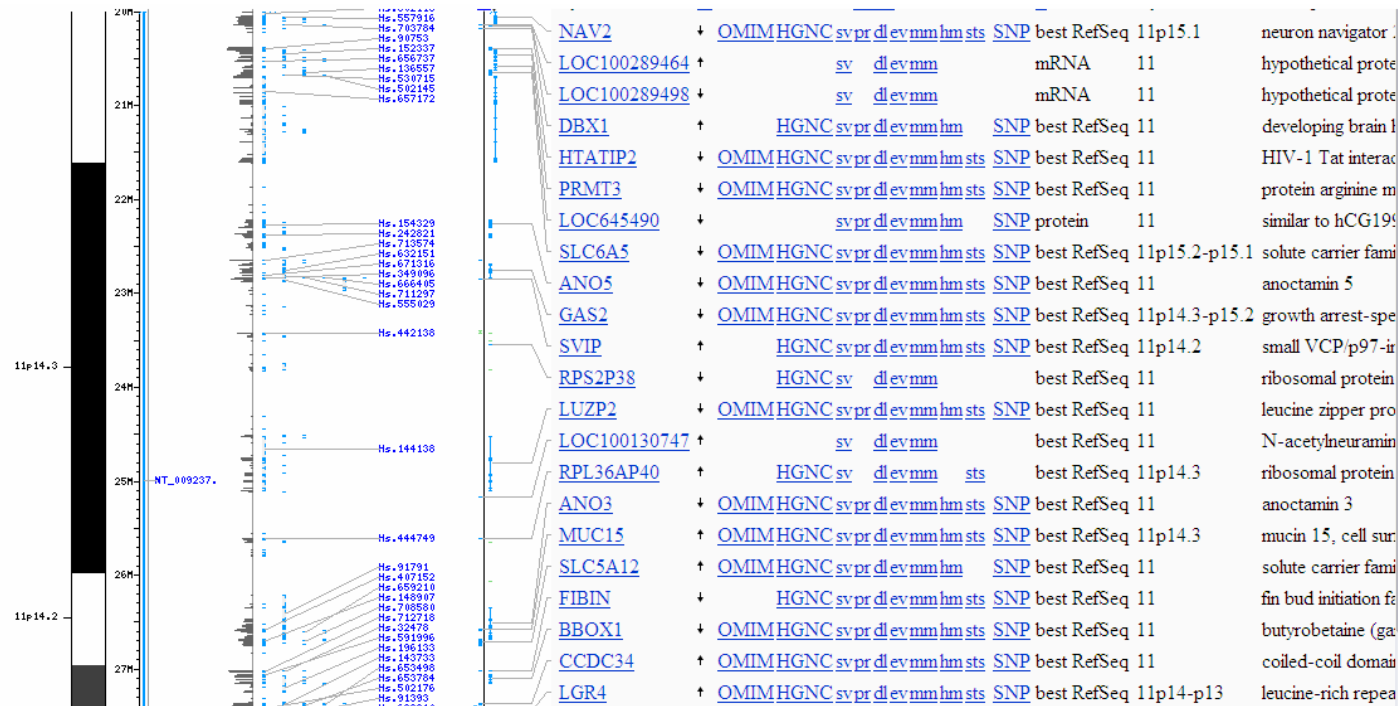


From: Strachan and Read  
Human Molecular Genetics

# Positional cloning

## [Step 2: define the genes within the interval]

- This is easy (in 2009): go to Ensembl, NCBI, UCSC's genomic browser, etc.



# Positional cloning

## [Step 3: prioritize the candidates]

- Ideally, one should “prioritize” the genes to be screened, among all those that are present in the identified interval
- The procedure is technically called “target prioritization” and is more or less the same as the one described previously for the candidate gene approach

# Positional cloning

## [Step 4: obtain the template DNA]

- To screen for mutations, it is necessary to obtain enough DNA from patients, also in a suitable form for sequencing. This is achieved by either:
  1. Library screening (old-fashioned, time consuming)
  2. Exon PCR
  3. Long-range PCR
  4. Microarray-based sequence capture (since 2008!)
  5. ...?

# Positional cloning

## [Step 4: obtain the template DNA]

### Library screening

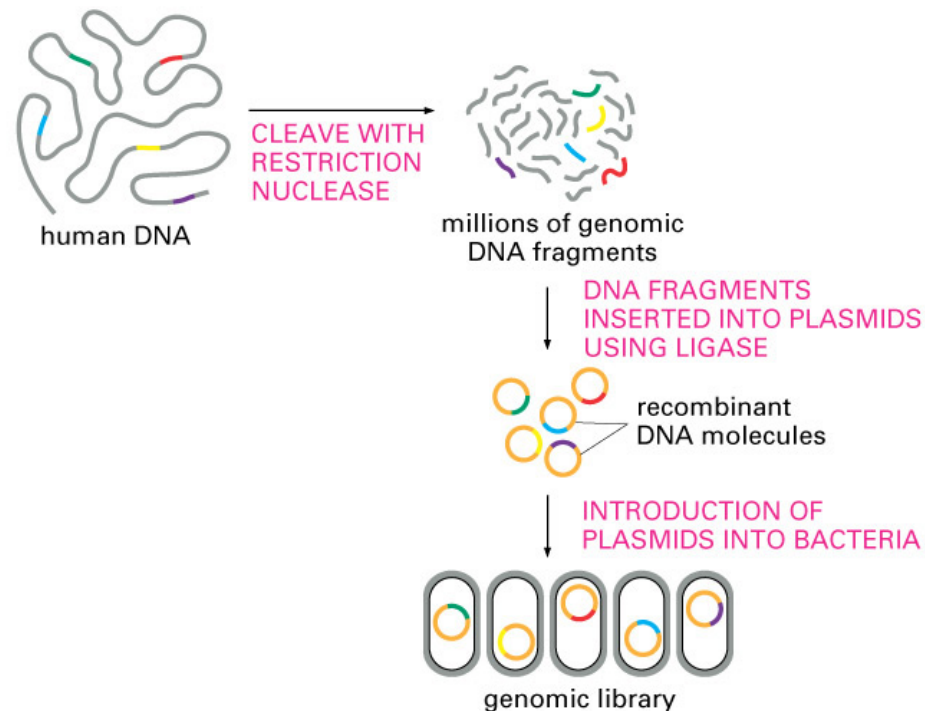


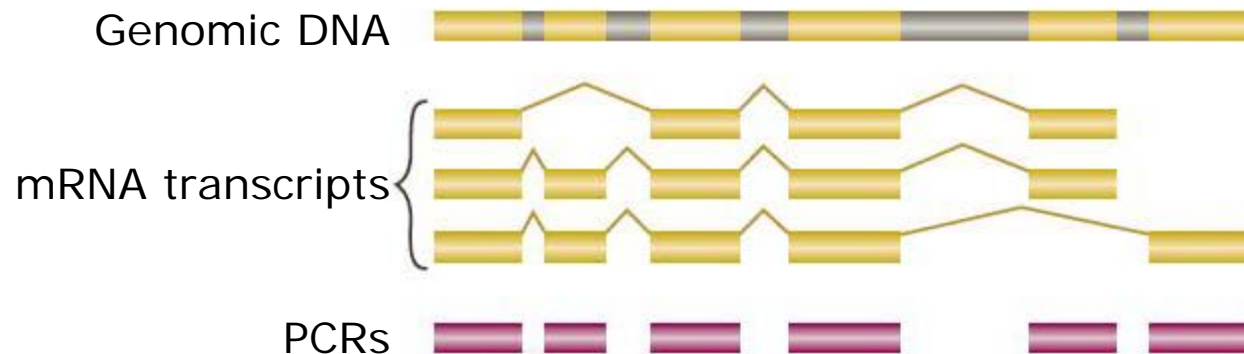
Figure 10-23 Essential Cell Biology, 2/e. (© 2004 Garland Science)

Limitations: **VERY** time-consuming

# Positional cloning

## [Step 4: obtain the template DNA]

### Exon PCR



Advantages: relatively quick processing

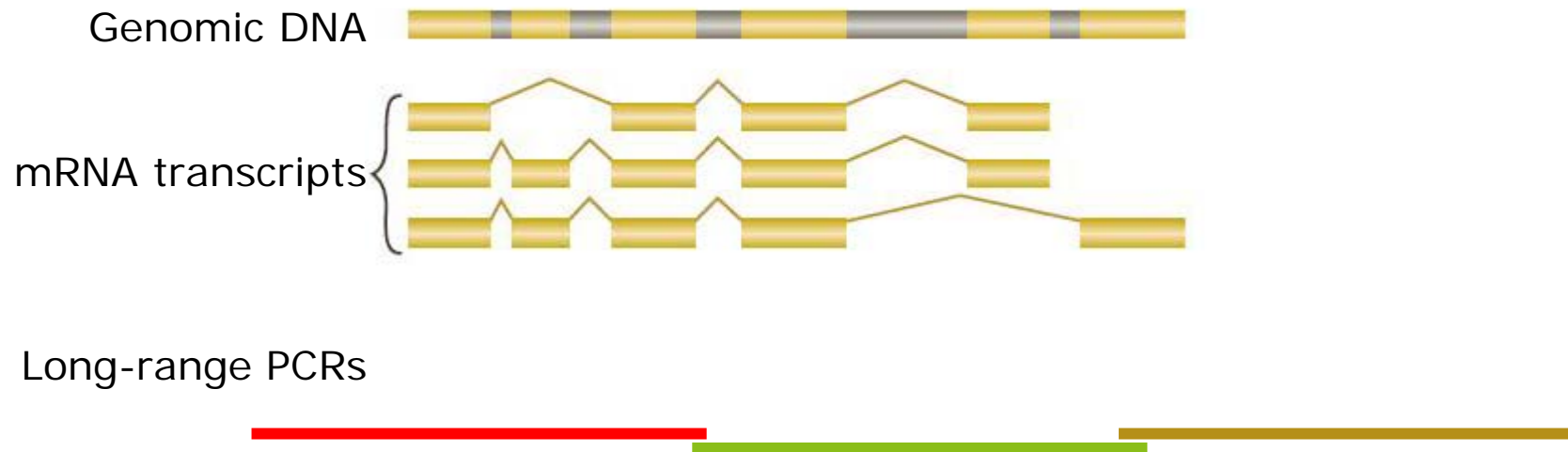
Limitations: only exons are screened



# Positional cloning

## [Step 4: obtain the template DNA]

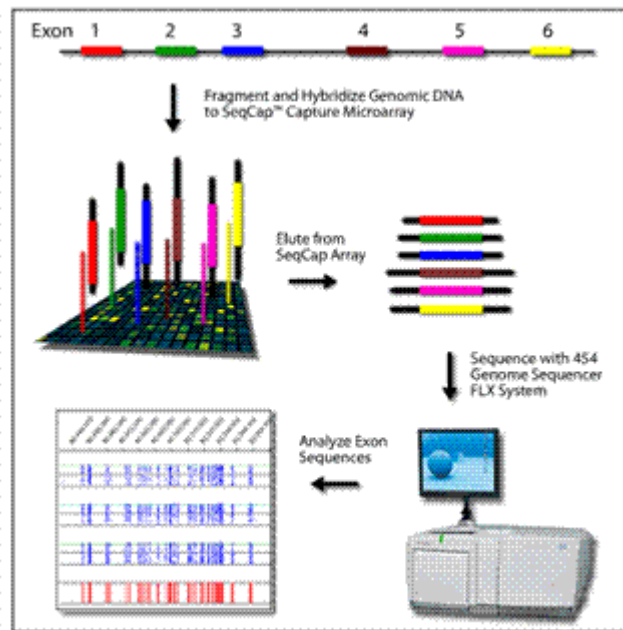
### Long-range PCR



# Positional cloning

## [Step 4: obtain the template DNA]

### Microarray-based sequence capture



Advantages: large-scale approach  
(hundreds of genes can be investigated)

Limitations: many false positives

# Positional cloning

## [Step 5: Find the DNA variants, etc.]

- Once it has been obtained, the template DNA is sequenced and the results analyzed as described previously for the candidate gene approach (cosegregation analyses, statistical test, functional tests, etc.)